

AD _____

Award Number: DAMD17-02-1-0484

TITLE: Gene Expression Analysis of Breast Cancer Progression

PRINCIPAL INVESTIGATOR: William L. Gerald, M.D., Ph.D.

CONTRACTING ORGANIZATION: Sloan-Kettering Institute for Cancer Research
New York, NY 10021

REPORT DATE: July 2005

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20050916 085

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE 01-07-2005		2. REPORT TYPE Final		3. DATES COVERED 17 Jun 2002 – 16 Jun 2005	
4. TITLE AND SUBTITLE Gene Expression Analysis of Breast Cancer Progression				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER DAMD17-02-1-0484	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) William L. Gerald, M.D., Ph.D.				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Sloan-Kettering Institute for Cancer Research New York, NY 10021				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Breast cancer (BC) is a heterogeneous disease with varying clinical behavior, and response to therapy that cannot be predicted based on existing classifications. It is the primary goal of our research to identify and characterize biological pathways and individual molecular components that play a primary role in BC development and progression. In order to identify genes, gene expression profiles and molecular pathways associated with metastatic BC we performed genome-wide gene expression analysis of a large number of breast cancer samples. Both unsupervised and supervised analyses were used to identify genes differentially expressed among samples and molecular subclasses of breast cancers. We identified a unique subclass of ER- breast carcinoma and characterized the molecular phenotype. In addition formal statistical testing was used to identify genes with marked changes in expression during progression. Lymph node metastases in particular showed significant decreases in the expression of many genes corresponding to extracellular matrix proteins and proteases when compared to matched primaries. Further expression changes in a variety of genes were associated with distant metastases. Immunohistochemistry and in situ hybridization were used to validate and extend findings. A variety of invitro and in vivo models have been used to elucidate specific molecular correlations.					
15. SUBJECT TERMS Breast Cancer					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code)
			UU	170	

Table of Contents

Cover.....	1
SF 298.....	2
Table of Contents.....	3
Introduction.....	4
Body.....	4
Key Research Accomplishments.....	4-5
Reportable Outcomes.....	5
Conclusions.....	6
References.....	6
Appendices.....	6
List of Personnel.....	
Bibliography.....	

INTRODUCTION

Breast cancer (BC) is a heterogeneous disease with varying clinical behavior and response to therapy that cannot be predicted based on existing clinical and pathologic classifications. This has led to an intense effort to understand the biology of BC and a search for genes and gene products that play a major role in tumor development and progression. A comprehensive analysis of gene expression can provide crucial clues concerning the intrinsic biology of a cancer and ultimately contribute to diagnostic decisions and therapies tailored to an individual patient. New, high-throughput mRNA analysis platforms, such as DNA microarrays, allow comprehensive measurement of gene expression and can produce large data sets with the potential to provide novel insights into biology at the molecular level. Our studies are designed to identify gene expression profiles that are associated with tumor progression and can be used for discrimination of clinically relevant subgroups of BC. An understanding of the mechanisms that drive progression of BC will provide biomarkers for diagnosis, risk stratification and therapeutic targets that could have an enormous impact on the care of these patients. The specific aims of our project are: 1) To identify the genes, gene expression profiles and molecular pathways associated with metastatic BC using microarray based, gene expression analysis and comparison of concurrent primary and metastatic tumors within the same patients. 2) To identify gene expression differences associated with clinical outcome by comparison of comprehensive expression profiles from stage and histology matched primary BCs in patients with long term recurrence-free survival and patients that die of metastatic disease.

BODY

We have completed all tasks originally proposed. Specifically we have identified and processed all tissue samples planned for specific aims 1 and 2. RNA has been isolated and labeled cRNA target from these samples has been subjected to gene expression analysis using oligonucleotides microarrays with features for over 33000 genes/ESTs. Hierarchical clustering of the gene expression data showed that most samples grouped according to estrogen receptor status (ER). In addition, the matched primary carcinomas and lymph node metastases have global expression profiles more similar to each other than to other breast cancers. Both unsupervised and supervised analyses were used to identify genes differentially expressed among samples and molecular subclasses of breast cancers. We identified a unique subclass of ER- breast carcinoma and characterized the molecular phenotype (Doane et al. appendix). In addition formal statistical testing was used to identify genes with marked changes in expression during progression. Lymph node metastases in particular showed significant decreases in the expression of many genes corresponding to extracellular matrix proteins and proteases when compared to matched primaries. Further expression changes in a variety of genes were associated with distant metastases. Immunohistochemistry and in situ hybridization were used to validate and extend findings. A variety of invitro and in vivo models have been used to elucidate specific molecular correlations (Dechow, et al. Bhargava et. al., Minn et al., Kang et al. appendix).

KEY RESEARCH ACCOMPLISHMENTS

- 1) Evaluation and selection of tumor cases to be used for specific aims 1 and 2
- 2) Microdissection of frozen tissue, RNA preparation and analysis of all samples.
- 3) Microarray based gene expression analysis of all samples.

- 4) Analysis of data from specific aims 1 and 2 and identification of differentially expressed genes.
- 5) Validation of differential expression at the RNA and protein level for select genes.
- 6) Identification of genes that participate in distinct organ-specific metastasis
- 7) Identification of a unique estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and proliferative response to androgen.

REPORTABLE OUTCOMES

Dechow TN, Pedranzini L, Leitch A, Leslie K, Gerald WL, Linkov I, Bromberg JF. Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C. *Proc Natl Acad Sci U S A*. 2004 101(29):10602-7.

Bhargava R, Gerald W, Lal P, and Chen B. Epidermal growth factor receptor (EGFR) gene amplification in breast cancer: Correlation with mRNA and protein expression and absence of common activating mutations. *Mod Pathol*. Epub ahead of press 2005.

Minn A, Kang Y, Serganova I, Gupta G, Giri D, Doubrovin M, Ponomarev V, Gerald W, Blasberg R, Massague J. Distinct organ-specific metastasis potential of individual breast cancer cells and primary tumors. *J Clin Invest*. 115: 44-55, 2005

Minn A, Gupta G, Siegel P, Bos P, Shu W, Giri D, Viale A, Olshen A, Gerald W, Massague J. Genes that predict and mediate breast cancer metastasis to the lung. In press *Nature*.

Kang Y, He W, Gupta G, Tulley W, Serganova I, Chen C, Manova-Todorova K, Blasberg R, Gerald W and Massagué J. The Smad4 Tumor Suppressor Mediates Pro-Metastatic TGF β Gene Responses in Breast Cancer Bone Metastasis. Submitted

Doane A, Danso M, Lal P, Donaton M, Zhang L, Hudis C, and Gerald W. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and proliferative response to androgen. Submitted

Donaton M, Giri D, Olshen A, Panageas K, Levcovici S, Lal P, Brogi E, Hudis C, VanZee K, Tan L, Gerald W *Comprehensive gene expression analysis of paired primary breast carcinomas and lymph node metastases*. Abstract presentation American Association of Cancer Research, 2003.

Giri D, Donaton M, Olshen A, Panageas K, Levcovici S, Lal P, Brogi E, Hudis C, VanZee K, Tan L, Gerald W. *Gene expression differences between paired primary and metastatic breast carcinomas*. Abstract presentation United States and Canadian Academy of Pathology, 2003.

Lal P, Donaton M, Giri D, Chen B, Gerald W Molecular Diagnosis of Breast Cancer Therapeutic Biomarkers Using Oligonucleotide Arrays Abstract presentation USCAP 2005.

Doane A, Danso M, Lal P, Donaton M, Zhang L, and Gerald W. Estrogen Receptor-Negative Breast Cancer with an Active Hormone Response Pathway: Therapeutic Implications. Abstract presentation AACR 2005

CONCLUSIONS

Comprehensive gene expression analysis of archived breast cancer samples is feasible. Molecular subgroups of breast carcinoma identified by gene expression analysis are strongly influenced by the ER status of the tumor. The gene expression profiles of paired primary and metastatic breast carcinomas are remarkably similar and the differences observed appear to reflect different microenvironments and tissue specific responses to tumor growth. Taken together, these results suggest that molecular features of breast carcinomas metastatic to lymph nodes are largely present in the primary tumor and might have been acquired early in tumorigenesis.

Analysis of primary tumors from patients with differing outcomes demonstrated a relatively small number of genes associated with progression. However several have interesting functional attributes that could impact on tumor biology. Functional analysis is providing strong evidence that some of these differentially expressed genes will provide clinically useful biomarkers and therapeutic targets.

REFERENCES

None

APPENDICES

Dechow TN, Pedranzini L, Leitch A, Leslie K, Gerald WL, Linkov I, Bromberg JF. Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C. *Proc Natl Acad Sci U S A*. 2004 101(29):10602-7.

Bhargava R, Gerald W, Lal P, and Chen B. Epidermal growth factor receptor (EGFR) gene amplification in breast cancer: Correlation with mRNA and protein expression and absence of common activating mutations. *Mod Pathol*. Epub ahead of press 2005.

Minn A, Kang Y, Serganova I, Gupta G, Giri D, Doubrovin M, Ponomarev V, Gerald W, Blasberg R, Massague J. Distinct organ-specific metastasis potential of individual breast cancer cells and primary tumors. *J Clin Invest*. 115: 44-55, 2005

Minn A, Gupta G, Siegel P, Bos P, Shu W, Giri D, Viale A, Olshen A, Gerald W, Massague J. Genes that predict and mediate breast cancer metastasis to the lung. In press *Nature*.

Yibin Kang, Wei He, Gaorav P. Gupta, Shaun Tulley, Inna Serganova, Chang-Rung Chen, Katia Manova-Todorova, Ronald Blasberg, William L. Gerald and Joan Massagué The Smad4 Tumor Suppressor Mediates Pro-Metastatic TGF β Gene Responses in Breast Cancer Bone Metastasis. Submitted

Doane A, Danso M, Lal P, Donaton M, Zhang L, Hudis C, and Gerald W. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and proliferative response to androgen. Submitted

Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C

Tobias N. Dechow^{*†}, Laura Pedranzini[†], Andrea Leitch[†], Kenneth Leslie[†], William L. Gerald[†], Irina Linkov[†], and Jacqueline F. Bromberg^{†‡}

^{*}Laboratory of Molecular Cell Biology, The Rockefeller University, New York, NY 10021; and Departments of [†]Medicine and [‡]Pathology, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021

Communicated by James E. Darnell, Jr., The Rockefeller University, New York, NY, June 9, 2004 (received for review December 2, 2003)

Persistently activated Stat3 is found in many different cancers, including ~60% of breast tumors. Here, we demonstrate that a constitutively activated Stat3 transforms immortalized human mammary epithelial cells and that this oncogenic event requires the activity of matrix metalloproteinase-9 (MMP-9). By immunohistochemical analysis, we observe a positive correlation between strong MMP-9 expression and tyrosine phosphorylated Stat3 in primary breast cancer specimens. These results demonstrate a relationship between activated Stat3 and MMP-9 in breast oncogenesis.

Signal transducer and activator of transcription (STAT) proteins are a family of transcription factors that are normally inactive within the cytoplasm of cells and become activated by tyrosine phosphorylation in response to cytokines and growth factors. Dimerization through reciprocal SH2-phospho-tyrosine interactions of tyrosine-phosphorylated STATs leads to their accumulation in the nucleus where they bind DNA and activate transcription. STAT dimers are dephosphorylated within the nucleus and transported back to the cytoplasm (1). In normal cells, STAT activation is transient whereas, in a large number of primary tumors and cancer-derived cell lines, STAT proteins (in particular Stat3) remain activated by persistently activated tyrosine kinases and/or a decrease in the negative regulators of STAT dephosphorylation (2). Introduction of dominant negative Stat3 or Stat3 antisense oligonucleotides leads to induction of apoptosis, decreased angiogenesis, or growth arrest of cancer-derived cell lines, including breast cancer cells (2, 3). In addition, a constitutively active mutant form of Stat3, Stat3-C, which is dimerized by cysteine-cysteine residues instead of pY-SH2 interactions, can transform immortalized cultured rodent fibroblasts (4). Stat3 is persistently tyrosine phosphorylated (by immunohistochemical and biochemical analyses) in 30–60% of primary breast cancer specimens (3, 5–7), leading us to test whether Stat3-C could mediate transformation of immortalized human mammary epithelial cell lines (HMECs), possibly more relevant to human tumor biology.

We report here that Stat3-C can transform immortalized HMECs and have determined that matrix metalloproteinase-9 (MMP-9) activity is increased in the Stat3-C-containing cell lines and that this activity is required for Stat3-C-mediated anchorage-independent growth.

Experimental Procedures

Cells and Growth Conditions. MCF-10A cells were obtained from the American Type Culture Collection (ATCC). Immortalized HMECs (referred to as HMLHT) and *Hras*V12-transformed HMLHT cells were obtained from R. A. Weinberg (Massachusetts Institute of Technology, Boston) (8). Stat3-C and v-src-expressing cells were generated by retroviral infection as described (9). Puromycin (2 μ g/ml) was added for selection. Cell proliferation was determined after 7 days by using alamarBlue (BioSource International, Camarillo, CA).

Plasmids and Reagents. pBabe-Stat3-C was generated by inserting a *Bam*HI site 3' of Stat3-C RcCMV (4) and subcloning the *Bam*HI cDNA insert into pBabe-Puro (10). PBabe-vsrc was from H. Hanafusa (Osaka Bioscience Institute, Osaka, Japan). The MMP-9 promoter luciferase pGL2 construct was obtained from M. Seiki (Kanazawa University, Kanazawa, Japan) (11). The MMP-2/9 inhibitor II (Calbiochem) was resuspended in DMSO (50 μ M) and subsequently diluted in PBS for further use. Recombinant MMP-9 was obtained from R & D Systems.

Soft Agar Assays. Soft agar assays were performed as described (8). HMLHT cells (2×10^4) and MCF-10A cells (2×10^5) were seeded per six-well in triplicate in 3 ml of top-agar. Colonies were stained with 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) (Sigma).

S.C. Tumorigenicity Assays. Six- to 8-week-old immunocompromised nonobese diabetic (NOD)/severe combined immunodeficient (SCID) mice (Taconic) were γ -irradiated with 300 rad, 4 h before injection to suppress natural killer cell activity. Cells (5×10^6) were harvested, mixed with an equal volume of Matrigel (Becton Dickinson), and injected in the mouse flank. Tumor size was measured once a week. Mice were killed after 10 weeks of observation or after the tumor grew to ≈ 600 mm³. Nuclear extracts were isolated from the tumors and analyzed for the presence of Stat3-C by anti-Flag Western blots.

Gene Array Analysis. For gene array analysis, see *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site.

RT-PCR for MMP-9. RT-PCR for MMP-9 was performed by preparation of total RNA with RNeasy (Qiagen, Valencia, CA) followed by RT (Clontech). PCR reactions were performed by using MMP-9 primers (5'-primer GATGCGTG-GAGAGTCGAAAT; 3'-primer CACCAAACTGGATGAC-GATG). GAPDH primers were used for loading control as described (12).

Western Blots, Immunoprecipitation, Zymography, Electrophoretic Mobility-Shift Assay (EMSA), Luciferase, MMP-9 Activity ELISA, and Immunocytochemistry. Cytoplasmic and nuclear extracts were prepared as described (4). Anti-Flag antibody (M2, Sigma) was diluted 1:1,000. MMP-9 antibody (Ab-2, Oncogene Research Products) was used for immunoprecipitations (1:20) and Western blots (1:1,000). Zymograms were performed as described

Abbreviations: STAT, signal transducer and activator of transcription; MMP-9, matrix metalloproteinase-9; HMEC, human mammary epithelial cells; EMSA, electrophoretic mobility-shift assay; APMA, 4-aminophenylmercuric acetate.

[†]To whom correspondence should be addressed. E-mail: bromberj@mskcc.org.

© 2004 by The National Academy of Sciences of the USA

(13, 14). EMSA was carried out as described by using a high-affinity m67 binding probe (4). HMLHT cells (2×10^4 /24-well dish) were transiently transfected with 0.4 μ g of MMP-9 Luciferase construct and 0.4 μ g of either pBabe or pBabe Stat3-C, by using Lipofectamine 2000 (GIBCO/BRL). Luciferase activity (Promega) was measured 24 h later. MMP-9 activity ELISA (Amersham Pharmacia) was conducted according to the manufacturer's instructions. *In situ* zymography was performed as described (15). HMLHT cells grown on multichamber slides were overlaid with DQ gelatin (100 μ g/ml) for 2 h at 37°C, washed, stained with 4',6-diamidino-2-phenylindole (DAPI), fixed, and analyzed by confocal laser microscopy. Immunocytochemistry was performed by fixing cells in 50:50 acetone:methanol and permeabilized with 0.1% Triton X-100. MMP-9 Ab-1 (Oncogene Research Products) was added overnight at 4°C (1:20).

Immunohistochemistry. Multitissue blocks of formalin-fixed, paraffin-embedded breast cancer tissue (containing four representative 0.6-mm cores) were prepared by using a tissue array, and immunohistochemistry was performed as described (5). Antigen retrieval using citric acid (pH 6.0) at 97°C for 30 min was followed by treatment with 3% H₂O₂. Phospho-Stat3 (Tyr-705) antibody (Cell Signaling Technology, Beverly, MA) was used at 1:200 dilution. The phospho-peptide used for generating the antibody was used to confirm specificity of antibody binding. MMP-9 antibody (NCL-MMP9, Novocastra, Newcastle, U.K.) was used at 1:50 dilution. Scoring of the tissue microarray was performed by two independent observers (J.F.B. and T.N.D.) with a high correlation between scorers ($P < 0.001$) for both pStat3 and MMP-9. In order for a tumor to be considered positive for either pStat3 or MMP-9, all four replicates in the tissue array had to have a similar staining intensity; otherwise it was excluded. Statistical analyses were done by using STATVIEW (SAS Inst., Cary, NC). The correlation between the scores of both scorers and the relationship between that of pStat3 and MMP-9 were measured by using the χ^2 test.

Results

Stat3-C Transforms HMEC Cell Lines. Given the incidence of phosphorylated Stat3 in primary breast cancer specimens, we wished to determine whether the introduction of a constitutively activated version of Stat3 (Stat3-C) was sufficient for mediating transformation of HMECs. For these studies, we used two different immortalized nontransformed HMEC lines. HMECs from reduction mammaplasties were immortalized by introducing both SV40 large-T antigen and the telomerase catalytic subunit (8). MCF-10As are a spontaneously immortalized human breast epithelial cell line mutant in the cdk inhibitor p16 (9). Immortalized HMECs (referred to as HMLHT cells in this article) and MCF-10A cell lines have many of the characteristics of normal breast epithelium and do not form tumors in nude mice nor form colonies in soft agar, but undergo transformation upon the introduction of Ha-ras (8, 16).

Flag-tagged Stat3-C was introduced into MCF-10A and HMLHT cells by retroviral gene transfer, and polyclonal populations were selected. Western blot analysis showed expression of Stat3-C in both MCF-10A and HMLHT cells (Fig. 1A). EMSA of extracts from Stat3-C-expressing cells showed strong binding to a high-affinity Stat3 binding site (m67) in contrast to extracts from cell lines harboring the empty retroviral vector (Fig. 1B). The DNA-protein complex could be supershifted with an anti-Flag antibody but not by an anti-Stat1 antibody (data not shown).

A classical assay for cellular transformation is anchorage-independent growth. Control and Stat3-C-expressing MCF-10A and HMLHT cells were plated in soft agar, and colony formation after 3 weeks by Stat3-C-expressing cell lines but not control lines was evident (Fig. 1C).

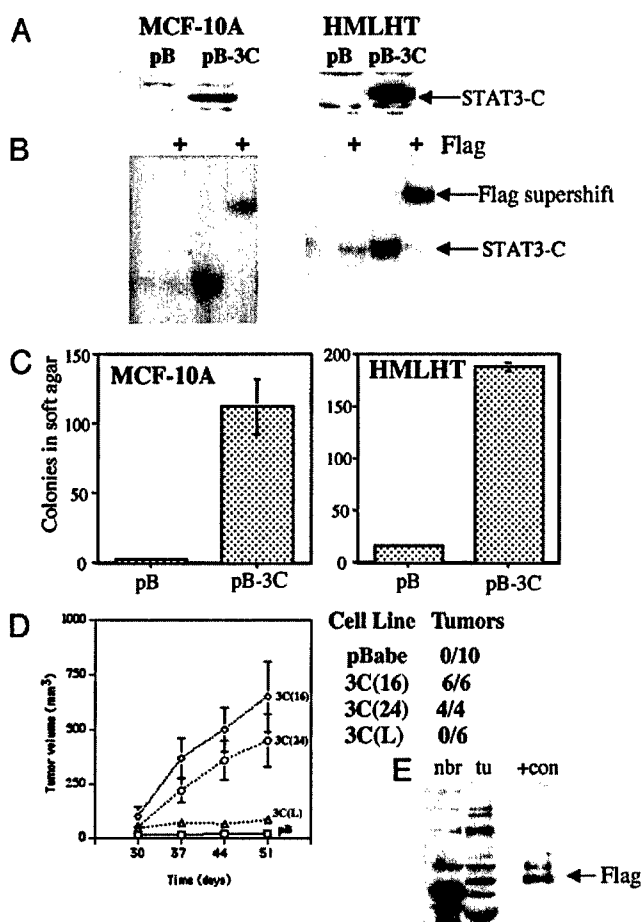


Fig. 1. Stat3-C induces tumorigenesis of HMLHT and MCF-10A cells in a dose-dependent manner. (A) Anti-Flag Western blot showing Stat3-C expression in MCF-10A and HMLHT cells expressing pBabe control vector (pB) and pBabe-Stat3-C (pB-3C). (B) EMSA performed with nuclear extracts from cell lines described in A. Stat3-C DNA binding was supershifted with anti-Flag antibody, indicated with a +. (C) Colony formation in soft agar of empty retroviral control (pB) and Stat3-C infected (pB-3C) MCF-10A and HMLHT cells (mean \pm SD). (D) Tumor growth in nonobese diabetic/severe combined immunodeficient (NOD/SCID) mice when using the HMLHT pBabe control cell line and subclones with high Stat3-C expression levels (no. 16 and no. 24) or low Stat3-C expression (3CL). Results are expressed as the mean of 4–10 tumors \pm SD at the indicated times after injection. (E) Nuclear extracts from a Stat3-C-derived tumor (tu), normal murine breast tissue (nbr), and cell line no. 16 (+con) were analyzed for the presence of Stat3-C by Flag immunoblot.

To determine whether the amount of Stat3-C expressed influenced the efficiency of transformation, single clones were isolated, and DNA-binding assays were carried out. Low (L) and high (H) Stat3-C-expressing clones were isolated and compared with the heterogeneous population (pB-3C) (see Fig. 6A, which is published as supporting information on the PNAS web site). Cells expressing low levels of Stat3-C did not grow in soft-agar, whereas higher expression levels (H) showed colony formation suggesting that a threshold amount of Stat3-C is required for soft-agar growth (see Fig. 6B).

Two high-expressing Stat3-C clones (no. 24 and no. 16) were injected s.c. into the flank of irradiated nonobese diabetic/severe combined immunodeficient (NOD/SCID) mice and gave rise to tumors in all animals in contrast to cells bearing the empty retroviral vector or a low-expressing clone (L) (Fig. 1D). The presence of Stat3-C within the tumor was determined by anti-Flag Western blot analysis (Fig. 1E). Thus, Stat3-C can mediate transformation of immortalized human breast epithelial cells.

This finding is an extension of our previous report that Stat3-C induced transformation of immortalized murine fibroblasts (4).

Stat3-C Induced Gene Expression. It is logical that the mechanism(s) by which this persistently active transcription factor mediates cellular transformation is through activation of specific genes. We next wished to identify differentially expressed mRNAs in Stat3-C-containing HMLHT and MCF-10A cells. By RT-PCR analysis of mRNA, Cyclin D1, Bcl-xL, myc, and vascular endothelial growth factor (VEGF), known target genes of activated Stat3 in fibroblasts, were not increased in the Stat3-C-expressing cell lines compared with those bearing the empty retroviral vector (data not shown). Thus, Affymetrix Gene Chip Analysis was performed on RNA isolated from HMLHT-Stat3-C and MCF-10A-Stat3-C cell lines compared with their respective, vector-infected control cells. One hundred and forty-one mRNAs were up-regulated, and 63 were down-regulated in the HMLHT-Stat3-C-expressing cells compared with HMLHT cells containing the empty retroviral vector; and 163 mRNAs were up-regulated and 36 were down-regulated in the MCF-10A-Stat3-C cells compared with MCF-10A cells bearing the empty vector (2-fold, $P < 0.001$). We then determined those mRNAs that were up- or down-regulated in both Stat3-C-expressing cell lines. Twenty-three mRNAs were increased, and one decreased in both cell lines (see Tables 1–3, which are published as supporting information on the PNAS web site). Some transcripts were increased by >8-fold in at least one of the Stat3-C-containing cell lines. However, the importance of these transcripts in tumorigenesis has not been well documented. One of the mRNAs up-regulated in both of the Stat3-C-expressing cell lines was MMP-9 (2.6- to 4-fold induction). Given the role of MMP-9 in tumor formation, invasion, metastasis, and angiogenesis (17), we focused our attention on this gene as possibly relevant to Stat3-C-mediated transformation in these breast epithelial cells.

MMP-9 Is Expressed and Zymographically Active in Stat3-C-Expressing HMEC Lines. Relative levels of MMP-9 mRNA were determined by RT-PCR in MCF-10A and HMLHT cells and found to be increased in the Stat3-C-expressing cells compared with empty retroviral vector-containing cells (Fig. 2A). To evaluate possible transcriptional regulation of MMP-9 by Stat3-C, we transiently transfected a luciferase construct containing the human MMP-9 promoter (with two potential Stat3-binding sites) with either empty vector or Stat3-C into HMLHT cells. Stat3-C expression led to a 4-fold increase of MMP-9 promoter-driven luciferase activity in HMLHT cells (Fig. 2B). MMP-9 (gelatinase B) is secreted as a 92-kDa pro-enzyme and cleaved by other proteases to an activated 84-kD form. By immunoprecipitation and Western blotting, latent MMP-9 protein was increased in the cell culture medium from Stat3-C-expressing HMLHT and MCF-10A cells compared with that in the medium from their respective control cell lines (Fig. 7A, which is published as supporting information on the PNAS web site). MMP-9 and MMP-2 (gelatinase B and A) are the two major gelatinases produced by cells. An increase in the latent 92-kDa MMP-9 was observed in the cell culture medium from Stat3-C-expressing cells compared with that from control-infected cells by gelatin zymography (Fig. 2C). The latent form of MMP-9 is active zymographically due to the denaturing conditions of SDS/PAGE, which reveals the catalytic domain of MMP-9. Notably, gelatin zymography did not reveal any 72-kDa, MMP-2 activity. Moreover, Stat3-C protein levels in HMLHT cells positively correlated with latent MMP-9 expression as determined by gelatin zymography (Fig. 7B). Thus, an increase of only the latent form of MMP-9 is observed in the cell culture medium of Stat3-C-expressing MCF-10A and HMLHT cells.

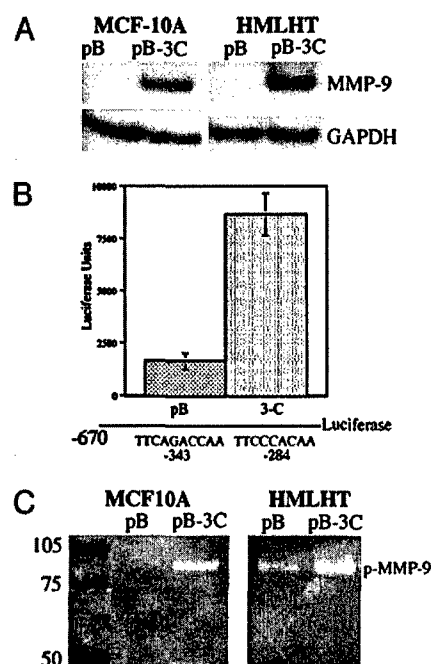


Fig. 2. Stat3-C-dependent induction of MMP-9 mRNA, luciferase activity, and protein. (A) Induction of MMP-9 mRNA in pBabe-Stat3-C (pB-3C)- and pBabe (pB)-infected MCF-10A and HMLHT cells determined by RT-PCR (Upper), normalized to GAPDH (Lower). (B) HMLHT cells were transfected with an MMP-9 promoter luciferase construct in conjunction with either pBabe (pB)- or pB Stat3-C (3-C)-expressing plasmids. Luciferase activities are shown as the mean \pm SD of three experiments performed in duplicate. (C) MMP-9 protein expression in cell culture medium from pBabe (pB)- and pBabe-Stat3-C (pB-3C)-infected MCF-10A and HMLHT cells shown by gelatin zymography.

Proteolytically Active MMP-9 Is Localized to the Cell Surface of Stat3-C-Containing Cells. A second assay for MMP-9 activity, which measures only cleaved (84-kD) protein, showed as expected no extracellular activity in either control or Stat3-C-expressing cells (Fig. 3A, black columns). In this assay, the total MMP-9 activity can be measured by treating the samples with 4-aminophenylmercuric acetate (APMA), which results in the cleavage of the MMP-9 pro-peptide, revealing enzymatically active MMP-9. After APMA treatment, an increase in MMP-9 in the medium from Stat3-C-expressing HMLHT cells was observed (Fig. 3A, gray columns). In contrast, total cell-associated MMP-9 activity was \approx 8-fold higher in Stat3-C-expressing HMLHT cells as compared with vector-infected cells (Fig. 3B, black columns). Treatment of these extracts with APMA led to only a modest increase in activity, suggesting that much of the cell-associated MMP-9 is in an enzymatically active form (Fig. 3B, gray columns). We also examined gelatinase activity *in situ* on cells grown in culture (Fig. 3C). Fluorescein-conjugated gelatin (DQ gelatin) was overlaid on cells, revealing an increase in fluorescence in the Stat3-C-expressing cells compared with control cells, which is a measure of the proteolytic activity of the gelatinase (Fig. 3C Upper). Furthermore, this activity was reduced in the presence of a dual specific MMP-2/9 enzymatic inhibitor, an N-sulfonylamino acid derivative that chelates zinc at the active site and inhibits MMP-2/9-dependent invasion, tumor growth, and metastasis in both cell culture and mouse tumor models (18, 19) (Fig. 3C Lower). Given the lack of MMP-2 expression in the Stat3-C-containing HMLHT cells as determined by zymography (Fig. 2C), we felt that this inhibitor was appropriate for the assay. The cellular localization of MMP-9 was examined by immunocytochemistry and was found to

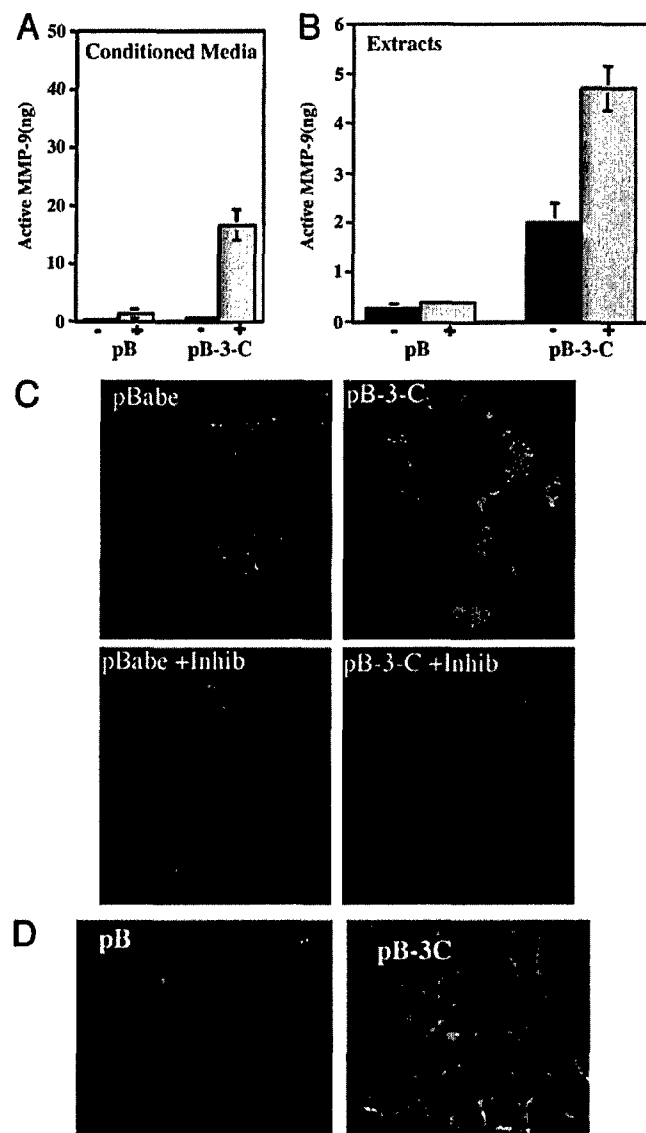


Fig. 3. Active MMP-9 is localized to the cell surface. (A and B) An ELISA specific for enzymatically active MMP-9 was performed on cell culture medium (A) and cell extracts (B) from pBabe (pB)- and pBabe-Stat3-C (pB-3C)-expressing cells. MMP-9 activity from cell culture medium and cell extracts was measured without (black columns) and with pretreatment with APMA (gray columns). Results are shown as the mean \pm SD of three experiments performed in duplicate. (C) *In situ* zymography of HMLHT cells expressing pBabe and pBabe-Stat3-C (pB-3-C) treated with DMSO (Upper) or 1.5 μ M MMP-2/9 inhibitor (Lower). The cells were then overlaid with DQ gelatin. Green staining indicates MMP-9-digested gelatin whereas blue indicates nuclear staining [4',6-diamidino-2-phenylindole (DAPI)]. (D) MMP-9 expression shown by immunofluorescence in the cell lines described in C.

be predominantly in a membrane-associated distribution (Fig. 3D).

Inhibition of MMP-9 Reduces Stat3-C-Dependent Transformation in HMLHT Cells. To determine whether the enzymatic activity of MMP-9 contributes to Stat3-C-induced anchorage-independent growth of HMLHT cells, a polyclonal population of Stat3-C-expressing cells and a high Stat3-C-expressing clone (data not shown) were grown in soft agar in the presence of the MMP-2/9 inhibitor. Colony formation was attenuated in the presence of increasing concentrations of the MMP-2/9 inhibitor (Fig. 4A).

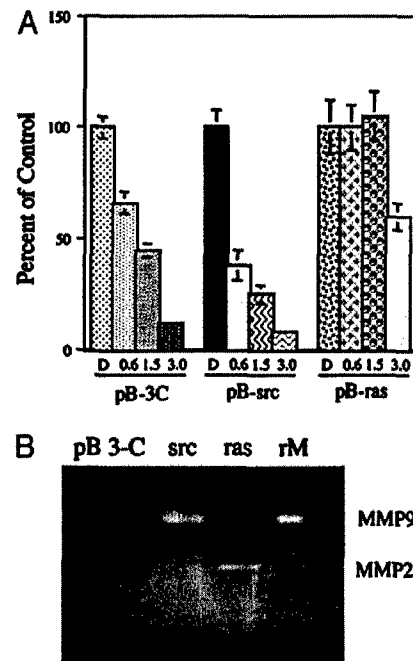


Fig. 4. MMP-9 activity is required for Stat3-C-dependent anchorage-independent growth. (A) Anchorage-independent growth of pBabe-Stat3-C cells (pB-3C)-, pBabe v-src (pB-src)-, and pBabe H-ras V12 (pB-ras)-expressing HMLHT cells. DMSO (D) control and increasing concentrations of MMP-2/9 inhibitor in μ M were added to the soft agar assay every other day (mean \pm SD). (B) Gelatin zymography of supernatants derived from HMLHT cells expressing either pBabe (pB), pBabe-Stat3-C (3C), pBabe v-src (src), or pBabe H-ras V12 (ras) and 0.5 ng of recombinant MMP-9 as a loading control.

The MMP-2/9 inhibitor did not influence the proliferation of Stat3-C-expressing HMLHT cells grown in monolayer culture (see Fig. 8, which is published as supporting information on the PNAS web site). Specificity of the MMP-2/9 inhibitor was examined in HMLHT cells transformed by either v-src or H-rasV12. Colony formation of HMLHT cells expressing v-src, an oncogene that activates Stat3 and requires Stat3 for its transforming capacity (20, 21), was suppressed by 1.5 μ M MMP-2/9 inhibitor (Fig. 4A). In contrast, H-rasV12-induced anchorage-independent growth of HMLHT cells was not affected by 1.5 μ M inhibitor (Fig. 4A). Gelatin zymography revealed high levels of latent MMP-9 in the medium of v-src-transformed HMLHT cells whereas the cell culture medium from H-rasV12-expressing cells did not have any detectable MMP-9 but did contain increased MMP-2 levels (Fig. 4B). These results demonstrate that MMP-9 activity is required for anchorage-independent growth of HMLHT cells induced by Stat3-C and v-src but not by H-rasV12.

MMP-9 Expression Correlates with That of Activated Stat3 in Primary Breast Cancer Specimens. Immunohistochemical analysis of microtissue arrays of primary human breast cancer specimens (34 tumor specimens and 8 normal) shows that 27% contain high levels (+++) of nuclear phospho-Stat3 (pStat3), 30% contain moderate levels of nuclear pStat3 (++), and 42% contain little to no pStat3 (0/+) (Fig. 5). Normal breast has little to no pStat3 (Fig. 5, Bottom). It has been determined that MMP-9 is over-expressed in primary breast carcinomas by immunohistochemistry (22–26). The cellular distribution of MMP-9 protein in paraffin sections is typically cytoplasmic (23–27). We stained sequential, serial sections of the breast microtissue arrays with anti-sera to MMP-9 and observed a strong cytoplasmic and perinuclear staining in 27% of these tumor specimens (+++),

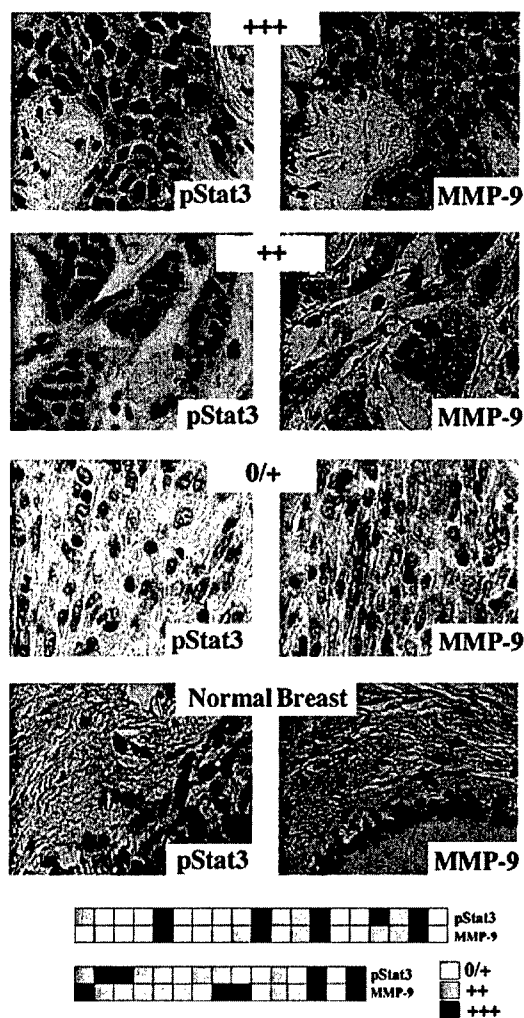


Fig. 5. Persistently phosphorylated Stat3 correlates with MMP-9 expression in primary breast cancer samples. Immunohistochemistry was performed on sequential sections of 34 primary breast cancer microtissue arrays with anti-phospho-Stat3 (pStat3) and anti-MMP-9 antibodies. A schematic overview of the tissue arrays and a summary of the immunohistochemistry results are shown. Representative sections of strong staining are indicated as +++ and shaded in black, moderate staining as ++ and shaded in gray, and weak to no staining as 0/+ and shaded in white. Normal breast had 0/+ staining for both pStat3 and MMP-9. A positive correlation was observed between (+++/++) staining for pStat3 and MMP-9 ($P < 0.001$) by χ^2 test.

moderate staining in 32% (++), and no to little staining in 38% (0/+) (Fig. 5). The majority of the MMP-9 staining was specific to the epithelial cells. However, the stromal cells surrounding the epithelial cells were also positive in two samples (data not shown). Not all samples that stained positively for pStat3 were also positive for MMP-9. However, a statistically positive correlation was observed between (+++/++) staining for pStat3 and MMP-9 ($P < 0.001$).

Discussion

Breast carcinogenesis is a process dependent upon the loss of tumor suppressors and gain of oncogenes. Our data suggest that activated Stat3 plays a role in breast tumorigenesis in part through the actions of MMP-9. Stat3 is persistently activated in a large fraction of primary breast cancers both by biochemical and immunohistochemical analyses (3, 5–7). Here, we demonstrate, by using two immortalized breast epithelial cell lines used to define oncogenes involved in breast tumorigenesis, the suffi-

ciency of Stat3-C in mediating transformation. We also determined that a threshold amount of Stat3-C is required for growth in soft agar and in nude mice.

Further characterization of Stat3-C-expressing MCF-10A and HMLHT cells did not reveal any significant differences in growth rate, growth-factor requirement, or resistance to proapoptotic stimuli (data not shown). The mechanism of transformation by Stat3-C is proposed to be through the genes it transcriptionally regulates. Some of the known targets of Stat3-C in fibroblasts were not altered in the breast epithelial cell lines. Transcriptional regulation of genes by activated Stat3 is likely dependent upon the cellular context and thus the mechanism of transformation. By Affymetrix Gene Chip analysis a short list of transcripts were identified (and many confirmed by RT/PCR) that were commonly up- or down-regulated in the Stat3-C-transformed cell lines (data not shown). Some of these transcripts may be involved in Stat3-C-mediated transformation, but we focused our attention on MMP-9.

By immunohistochemistry of cancer specimens, MMPs and in particular gelatinases have been found to be up-regulated in almost every tumor entity, including breast cancer (22–28). Cell culture and mouse experiments with mammary epithelial cells and cancer cells have revealed a crucial role for MMP-9 in tumor growth, invasion, metastasis, and angiogenesis (29–32). Many molecules and signaling pathways have been reported to be involved in the induction of MMP-9 in breast cancer cells, such as heregulin, estrogen, epidermal growth factor (EGF), c-jun, NF- κ B, and mitogen-activated protein kinase (MAPK) (28, 33–37). In addition to tumor-derived MMP expression, it is largely accepted that the tumor environment plays a crucial role in the activity of MMPs (17). Nevertheless, it has been demonstrated that expression of MMP-3/Stromelysin-1 is sufficient to transform mammary epithelial cells in culture as well as in a breast-specific transgenic mouse model, demonstrating an oncogenic potential of MMPs produced by epithelial cells (39).

Here, we show that MMP-9 mRNA and protein can be induced by Stat3-C in mammary epithelial cells. The MMP-9 promoter contains multiple putative Stat3-binding sites, two of which can be considered as high-affinity binding sites (11). However, a direct association between Stat3 and the MMP-9 promoter by chromatin immunoprecipitation has not been observed (data not shown). Nevertheless, an MMP-9 promoter luciferase construct (-670) is induced at least 4-fold by Stat3-C when transfected into HMLHT cells. We observed an increase in the levels of latent MMP-9 protein from conditioned media isolated from cells expressing Stat3-C. Furthermore, we demonstrated that proteolytically active MMP-9 is localized primarily to the cell surface, which is in accordance with prior studies supporting a role for cell surface-associated MMP-9 with respect to its enzymatic and biological activity (13, 14, 39). By using a dual-specific MMP-2/9 inhibitor, we observed suppression of anchorage-independent growth of Stat3-C and *v-src* (an oncogene that activates and requires Stat3 for transformation)-expressing cells but not of *H-ras*V12-transformed HMLHT cells. Thus, this inhibitor does not decrease growth in soft agar nonspecifically and indicates a crucial role for MMP-9 in anchorage-independent growth by Stat3-C and *v-src* in HMLHT cells.

We have examined the abundance and distribution of tyrosine-phosphorylated Stat3 in primary breast cancer samples and find that $\approx 30\%$ of the invasive tumors have strong staining for nuclear tyrosine phosphorylated Stat3. We did not have access to prognostic information with our tissue-array samples and therefore cannot say whether strong nuclear phospho-Stat3 is associated with indolent or aggressive breast cancer. Interestingly, high MMP-9 protein levels in sequential sections of the tissue micro arrays correlates with that of activated Stat3, supporting our cell culture work that MMP-9 induced by Stat3 may contribute to mammary tumorigenesis.

We thank James Darnell for helpful discussions, Robert Weinberg for cell lines, and Agnes Viale for assistance with Affymetrix analysis. This work was supported by National Institutes of Health Grant R01 CA87637, Department of Defense Concept Award BC996273, the

Speaker's Fund for Biomedical Research, a Charles E. Culpeper Scholarship Award, a Lerner Research Award (to J.F.B.), and a Deutsche Krebshilfe eingetragener Verein postdoctoral fellowship (to T.N.D.).

- Darnell, J. E., Jr. (1997) *Science* **277**, 1630–1635.
- Bromberg, J. (2002) *J. Clin. Invest.* **109**, 1139–1142.
- Garcia, R., Bowman, T. L., Niu, G., Yu, H., Minton, S., Muro-Cacho, C. A., Cox, C. E., Falcone, R., Fairclough, R., Parsons, S., et al. (2001) *Oncogene* **20**, 2499–2513.
- Bromberg, J., Wrzeszczynska, M., Devgan, G., Zhao, Y., Albanese, C., Pestell, R. & Darnell, J. E. J. (1999) *Cell* **98**, 295–303.
- Dolled-Filhart, M., Camp, R. L., Kowalski, D. P., Smith, B. L. & Rimm, D. L. (2003) *Clin. Cancer Res.* **9**, 594–600.
- Widschwendter, A., Tonko-Geymayer, S., Welte, T., Daxenbichler, G., Marth, C. & Doppler, W. (2002) *Clin. Cancer Res.* **8**, 3065–3074.
- Watson, C. J. & Miller, W. R. (1995) *Br. J. Cancer* **71**, 840–844.
- Elenbaas, B., Spirio, L., Koerner, F., Fleming, M. D., Zimonjic, D. B., Donaher, J. L., Popescu, N. C., Hahn, W. C. & Weinberg, R. A. (2001) *Genes Dev.* **15**, 50–65.
- Soule, H. D., Maloney, T. M., Wolman, S. R., Peterson, W. D., Jr., Brenz, R., McGrath, C. M., Russo, J., Pauley, R. J., Jones, R. F. & Brooks, S. C. (1990) *Cancer Res.* **50**, 6075–6086.
- Morgenstern, J. P. & Land, H. (1990) *Nucleic Acids Res.* **18**, 3587–3596.
- Sato, H. & Seiki, M. (1993) *Oncogene* **8**, 395–405.
- Yang, E., Wen, Z., Haspel, R. L., Zhang, J. J. & Darnell, J. E., Jr. (1999) *Mol. Cell. Biol.* **19**, 5106–5112.
- Fiore, E., Fusco, C., Romero, P. & Stamenkovic, I. (2002) *Oncogene* **21**, 5213–5223.
- Yu, Q. & Stamenkovic, I. (1999) *Genes Dev.* **13**, 35–48.
- Mira, E., Lacalle, R. A., Buesa, J. M., Gonzalez de Buitrago, G., Jimenez-Baranda, S., Gomez-Mouton, C., Martinez, A. C. & Manes, S. (2003) *J. Cell Sci.* **117**, 1847–1856.
- Ciardello, F., Gottardis, M., Basolo, F., Pepe, S., Normanno, N., Dickson, R. B., Bianco, A. R. & Salomon, D. S. (1992) *Mol. Carcinog.* **6**, 43–52.
- Egeblad, M. & Werb, Z. (2002) *Nat. Rev. Cancer* **2**, 161–174.
- Wroblewski, L. E., Pritchard, D. M., Carter, S. & Varro, A. (2002) *Biochem. J.* **365**, 873–879.
- Tamura, Y., Watanabe, F., Nakatani, T., Yasui, K., Fujii, M., Komurasaki, T., Tsuzuki, H., Maekawa, R., Yoshioka, T., Kawada, K., et al. (1998) *J. Med. Chem.* **41**, 640–649.
- Bromberg, J. F., Horvath, C. M., Besser, D., Lathem, W. W. & Darnell, J. E., Jr. (1998) *Mol. Cell. Biol.* **5**, 2553–2558.
- Turkson, J., Bowman, T., Garcia, R., Caldenhoven, E., De Groot, R. P. & Jove, R. (1998) *Mol. Cell. Biol.* **18**, 2545–2552.
- Giannelli, G., Fransvea, E., Marinosci, F., Bergamini, C., Daniele, A., Colucci, S., Paradiso, A., Quaranta, M. & Antonaci, S. (2002) *Biochem. Biophys. Res. Commun.* **292**, 161–166.
- Hanemaaijer, R., Verheijen, J. H., Maguire, T. M., Visser, H., Toet, K., McDermott, E., O'Higgins, N. & Duffy, M. J. (2000) *Int. J. Cancer* **86**, 204–207.
- Iwata, H., Kobayashi, S., Iwase, H., Masaoka, A., Fujimoto, N. & Okada, Y. (1996) *Jpn. J. Cancer Res.* **87**, 602–611.
- Jones, J. L., Glynn, P. & Walker, R. A. (1999) *J. Pathol.* **189**, 161–168.
- Scorilas, A., Karameris, A., Arnogiannaki, N., Ardavanis, A., Bassilopoulos, P., Tringas, T. & Talieri, M. (2001) *Br. J. Cancer* **84**, 1488–1496.
- Bodey, B., Bodey, B., Jr., Siegel, S. E. & Kaiser, H. E. (2001) *Anticancer Res.* **21**, 2021–2028.
- Kondapaka, S. B., Fridman, R. & Reddy, K. B. (1997) *Int. J. Cancer* **70**, 722–726.
- Weber, M. H., Lee, J. & Orr, F. W. (2002) *Int. J. Oncol.* **20**, 299–303.
- Li, H., Lindenmeyer, F., Grenet, C., Opolon, P., Menashi, S., Soria, C., Yeh, P., Perricaudet, M. & Lu, H. (2001) *Hum. Gene Ther.* **12**, 515–526.
- Mira, E., Manes, S., Lacalle, R. A., Marquez, G. & Martinez, A. C. (1999) *Endocrinology* **140**, 1657–1664.
- Yu, Q. & Stamenkovic, I. (2000) *Genes Dev.* **14**, 163–176.
- Reddy, K. B., Krueger, J. S., Kondapaka, S. B. & Diglio, C. A. (1999) *Int. J. Cancer* **82**, 268–273.
- Ricca, A., Biroccio, A., Del Bufalo, D., Mackay, A. R., Santoni, A. & Cippitelli, M. (2000) *Int. J. Cancer* **86**, 188–196.
- Tsai, M. S., Shamon-Taylor, L. A., Mehmi, I., Tang, C. K. & Lupu, R. (2003) *Oncogene* **22**, 761–768.
- Smith, L. M., Wise, S. C., Hendricks, D. T., Sabichi, A. L., Bos, T., Reddy, P., Brown, P. H. & Birrer, M. J. (1999) *Oncogene* **18**, 6063–6070.
- Razandi, M., Pedram, A., Park, S. T. & Levin, E. R. (2003) *J. Biol. Chem.* **278**, 2701–2712.
- Sternlicht, M. D., Lochter, A., Sympon, C. J., Huey, B., Rougier, J. P., Gray, J. W., Pinkel, D., Bissell, M. J. & Werb, Z. (1999) *Cell* **98**, 137–146.
- Stamenkovic, I. (2000) *Semin. Cancer Biol.* **10**, 415–433.

EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations

Rohit Bhargava, William L Gerald, Allan R Li, Qiulu Pan, Priti Lal, Marc Ladanyi and Beiyun Chen

Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

The human epidermal growth factor receptor (HER) family of receptor tyrosine kinase has been extensively studied in breast cancer; however, systematic studies of *EGFR* gene amplification and protein overexpression in breast carcinoma are lacking. We studied *EGFR* gene amplification by chromogenic *in situ* hybridization (CISH) and protein expression by immunohistochemistry in 175 breast carcinomas, using tissue microarrays. Tumors with >5 *EGFR* gene copies per nucleus were interpreted as positive for gene amplification. Protein overexpression was scored according to standardized criteria originally developed for HER-2. *EGFR* mRNA levels, as measured by Affymetrix U133 Gene Chip microarray hybridization, were available in 63 of these tumors. *HER-2* gene amplification by fluorescence *in situ* hybridization (FISH) and protein overexpression by immunohistochemistry were also studied. *EGFR* gene amplification (copy number range: 7–18; median: 12) was detected in 11/175 (6%) tumors, and protein overexpression was found in 13/175 (7%) tumors. Of the 11 tumors, 10 (91%) with gene amplification also showed *EGFR* protein overexpression (2+ or 3+ by immunohistochemistry). The *EGFR* mRNA level, based on Affymetrix U133 chip hybridization data, was increased relative to other breast cancer samples in three of the five tumors showing gene amplification. Exons 19 and 21 of *EGFR*, the sites of hotspot mutations in lung adenocarcinomas, were screened in the 11 *EGFR*-amplified tumors but no mutations were found. Three of these 11 tumors also showed *HER-2* overexpression and gene amplification. Approximately 6% of breast carcinomas show *EGFR* amplification with *EGFR* protein overexpression and may be candidates for trials of *EGFR*-targeted antibodies or small inhibitory molecules.

Modern Pathology advance online publication, 13 May 2005; doi:10.1038/modpathol.3800438

Keywords: breast cancer; *EGFR*; gene amplification; mRNA expression; mutation; protein overexpression; tissue microarray

The epidermal growth factor receptor (*EGFR*, *HER-1*, *c-erbB-1*) is one of the four transmembrane growth factor receptor proteins that share similarities in structure and function. Together, this group comprises the human epidermal growth factor receptor (*HER*) (*c-erbB*) family of receptor tyrosine kinases. The *EGFR* gene is located on the short arm of chromosome 7 and encodes a 170 kDa transmem-

brane protein consisting of an extracellular EGF-binding domain, a short transmembrane region, and an intracellular domain with ligand-activated tyrosine kinase activity.¹ Two ligands can activate *EGFR*: epidermal growth factor (*EGF*) and transforming growth factor- α (*TGF- α*). Ligand binding to *EGFR* results in receptor homo- or hetero-dimerization (with one of the *HER* family of receptor tyrosine kinases) followed by autophosphorylation of the tyrosine kinase domain.² Phosphorylated tyrosine residues serve as binding sites for the recruitment of signal transducers and activators of intracellular substrates. The Ras–Raf mitogen-activated protein kinase pathway and the phosphatidyl inositol 3' kinase and Akt pathway are the major signaling

Correspondence: Dr B Chen, MD, PhD, Department of Pathology, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA.

E-mail: chenb@mskcc.org

Received 17 February 2005; revised and accepted 8 April 2005; published online 13 May 2005

routes for the HER family, including EGFR.³⁻⁶ These pathways control several important biologic processes, including cellular proliferation, angiogenesis and inhibition of apoptosis.⁷

The interest in EGFR is further enhanced by the availability and FDA approval of specific EGFR tyrosine kinase inhibitors (eg, gefitinib). Many of these studies have focused on lung cancer, where approximately 10% of patients have a rapid and often dramatic clinical response.⁸⁻¹⁰ These gefitinib-responsive lung cancers have been found to contain somatic mutations in the tyrosine kinase domain of the *EGFR* gene.⁸⁻¹⁰ The data regarding the presence or absence of *EGFR* gene amplification in other tumor types, and their response to these EGFR tyrosine kinase inhibitors are still limited. EGFR protein overexpression has been reported to occur in 16-36% of breast cancers; however, systematic studies evaluating gene amplification, mRNA expression and protein expression in the same set of cases are lacking.¹¹⁻¹³ In order to address this issue, we studied 175 breast cancers for the presence of *EGFR* gene amplification. In addition, we analyzed EGFR protein expression, HER-2 protein expression and gene amplification in these tumors. We also examined *EGFR* transcript levels in a subset of these tumors by Affymetrix U133 chip hybridization and performed a mutational screen of the *EGFR*-amplified cases.

Materials and methods

Case Selection and Tissue Microarray Construction

In all, 188 randomly selected invasive breast carcinomas were included in this study. Tissue microarrays were created using 0.6 mm tissue cores as previously described.¹⁴⁻¹⁸ An H&E-stained section was evaluated for the presence of invasive breast carcinoma and the area to be used for creation of the tissue microarrays was marked on the slide and the donor block. Three to four cores from different areas of the tumor were sampled for each tumor.

Histologic Examination

Histologic assessment of tumor type and grade were routinely performed on 4-5 μ m thick H&E sections of formalin-fixed paraffin-embedded tumors. The nuclear grades of invasive ductal and lobular carcinomas were designated as follows: grade 1, small, regular uniform cells; grade 2, moderate increase in size and variability; grade 3, marked variation in size and shape. The architectural grades of invasive ductal carcinomas were designated as follows: grade 1, well developed (>75%) tubule formation; grade 2, moderate (10-75%) tubule formation; grade 3, little or no (<10%) tubule formation.

Immunohistochemistry

Tissue microarray sections (4-5 μ m thick) were used for all immunohistochemical analyses. The Ventana CONFIRMTM antiestrogen receptor (clone 6F11) and antiprogesterone receptor (clone 16) monoclonal antibodies were used for immunohistochemical analyses of estrogen receptor and progesterone receptor, respectively, performed on the Ventana automated slide stainers according to the manufacturer's instructions (Ventana Inc., Tucson, AZ, USA). The estrogen receptor or progesterone receptor results were manually screened and were interpreted as positive when more than 10% of tumor cells showed positive nuclear staining. HER-2 immunohistochemistry was performed using the HercepTestTM kit (DAKO Corp, Carpinteria, CA, USA) and EGFR immunohistochemistry was performed using a monoclonal EGFR antibody (Clone 31G7, Zymed Laboratories Inc., South San Francisco, CA, USA) according to the manufacturer's instructions; both HER-2 and EGFR results were interpreted manually as follows: 0, no membrane staining; 1+, faint, partial membrane staining; 2+, weak, complete membrane staining in >10% of invasive cancer cells; 3+, intense complete membrane staining in >10% of invasive cancer cells. The highest immunohistochemical score obtained among different cores of the same tumor was used as the final immunohistochemical result of that tumor.

Chromogenic *In Situ* Hybridization

Chromogenic *in situ* hybridization (CISH) for *EGFR* gene was performed according to the manufacturer's instructions. Briefly, the tissue microarray sections were incubated at 55°C overnight. The slides were deparaffinized in xylene and graded ethanol. Heat pretreatment was carried out in the pretreatment buffer (Zymed Laboratories Inc.) at 98-100°C for 15 min. The tissue was digested with pepsin for 10 min at room temperature. After application of Zymed SpotLight[®] digoxigenin labeled *EGFR* probe (Zymed Laboratories Inc.), the slides were coverslipped and edges sealed with rubber cement. The slides were heated at 95°C for 5 min followed by overnight incubation at 37°C using a moisturized chamber. Posthybridization wash was performed the next day and followed by immunodetection using the CISHTM polymer detection kit (Zymed Laboratories Inc.). The CISH signals were counted in at least 30 nuclei with a light microscope using a $\times 40$ objective. A tumor was interpreted as positive for gene amplification when the average number of gene copies was >5 per nucleus.

Fluorescence *In Situ* Hybridization

Fluorescence *in situ* hybridization (FISH) for *HER-2* was performed using the PathVysion HER-2 probe

kit (Vysis Inc. Downers Grove, IL, USA) as previously described.¹⁷ The signal enumeration was performed under $\times 1000$ magnification. The number of chromosome 17 signals, HER-2 signals, and number of tumor nuclei scored were recorded for each core. At least 30 cells were counted per tissue core. Tumors were interpreted as amplified when the ratio of HER-2/chromosome17 signals was ≥ 2.0 . The average ratio of different cores from the same tumor was used as the final score for determination of gene amplification status of that particular tumor.

EGFR mRNA Expression

EGFR mRNA levels were determined in a subset of cases using Affymetrix human genome U133 GeneChip[®] expression arrays. RNA extraction, RNA target synthesis, and target labeling were performed as previously described.¹⁹ Gene expression analysis was carried out using the Affymetrix U133A human gene array, which has 22 283 features for individual gene/EST clusters, using instruments and protocols recommended by the manufacturer. For each gene on every sample we extracted two response measures, the Average Difference and Absolute Call, as determined by the default settings of Affymetrix Microarray Suite 5.0. Expression values on each array were multiplicatively scaled to have an average expression of 500 across the central 96% of all genes on the array. Calculations of relative EGFR transcript levels were based on data from Affymetrix probe set 201984_s_at.

EGFR Mutation Analysis

Selected cases were analyzed for the presence of hotspot mutations in exon 19 (short in-frame deletions) and exon 21 (L858R mutation) that together account for approximately 90% of EGFR mutations detected in lung cancers.⁸⁻¹⁰ Exon 19 deletions were studied by length analysis of fluorescently labeled polymerase chain reaction (PCR) products on a capillary electrophoresis device, and the exon 21 L585R mutation was detected by PCR followed by Sau96I restriction enzyme digestion, based on a new Sau96I site created by the L585R mutation (2819T>G), followed by capillary electrophoresis of the Sau96I-digested fluorescently labeled PCR products. These sensitive assays can detect mutations in the presence of up to 90% non-neoplastic cells and are described in detail elsewhere.²⁰

Results

We obtained both CISH and immunohistochemistry EGFR data on 175 of the 188 breast cancers. Nine tumors failed both CISH and immunohistochemistry, four additional tumors failed immunohisto-

chemistry alone. The reasons for failure were a complete loss of tissue cores from the tissue microarrays, less than 30 tumor cells available for scoring, and absence of hybridization signals. The absence of signals probably resulted from under- or over-digestion since tissue digestion for a particular tumor cannot be adjusted on a tissue microarray.

EGFR gene copy number ranged from 2 to 18 in the samples studied. Copy number greater than 5 was considered amplified and identified in 11/175 (6%) tumors (Table 1). The gene copy number in amplified tumors ranged from 7 to 18 (mean: 12.1; median: 12) and in nonamplified tumors ranged from 2 to 5 (mean: 2.4; median: 2) (Figure 1). Affymetrix U133A data on mRNA levels for EGFR were available in five of the amplified cases. Three of these (Table 2) showed increased EGFR mRNA levels greater than two-fold of the average EGFR mRNA level in EGFR-nonamplified tumors, and the remaining two tumors showed no significant increase above the average EGFR mRNA level. The mRNA data were not available in the other six EGFR-amplified tumors. No statistically significant correlation between gene copy number and level of EGFR transcript was found in this small number of amplified cases. Of the 164 tumors without EGFR gene amplification, mRNA data were available in 56 tumors. All but one tumor showed normal mRNA levels. The discordant case showed a 7.4-fold increase in mRNA level (data not shown).

By immunohistochemistry, the majority of breast carcinomas demonstrated 0–1+ immunoreactivity (162/175, 94%). Eight of the 11 breast carcinomas with amplified EGFR showed 3+ immunoreactivity, two tumors demonstrated 2+ and one tumor was scored as 1+ (Table 1). There was a strong correlation between 3+ immunoreactivity and gene amplification ($P < 0.0001$, Fisher's exact test). Three of the 164 nonamplified tumors demonstrated EGFR protein overexpression. Two of these three tumors were poorly differentiated invasive ductal carcinomas and were 2+ by immunohistochemistry, the third tumor was an invasive pleomorphic lobular carcinoma and showed immunoreactivity of 3+ for EGFR without gene amplification.

Table 1 Correlation of EGFR gene amplification and protein expression

Immunohistochemistry	Gene amplification	No gene amplification	Total
0	0	151	151
1+	1 (9%)	10	11
2+	2 (50%)	2	4
3+	8 (89%*)	1	9
Total	11 (6%)	164	175

* $P < 0.0001$ (Fisher's exact test for EGFR immunohistochemistry 0–2+ and 3+ vs amplification status).

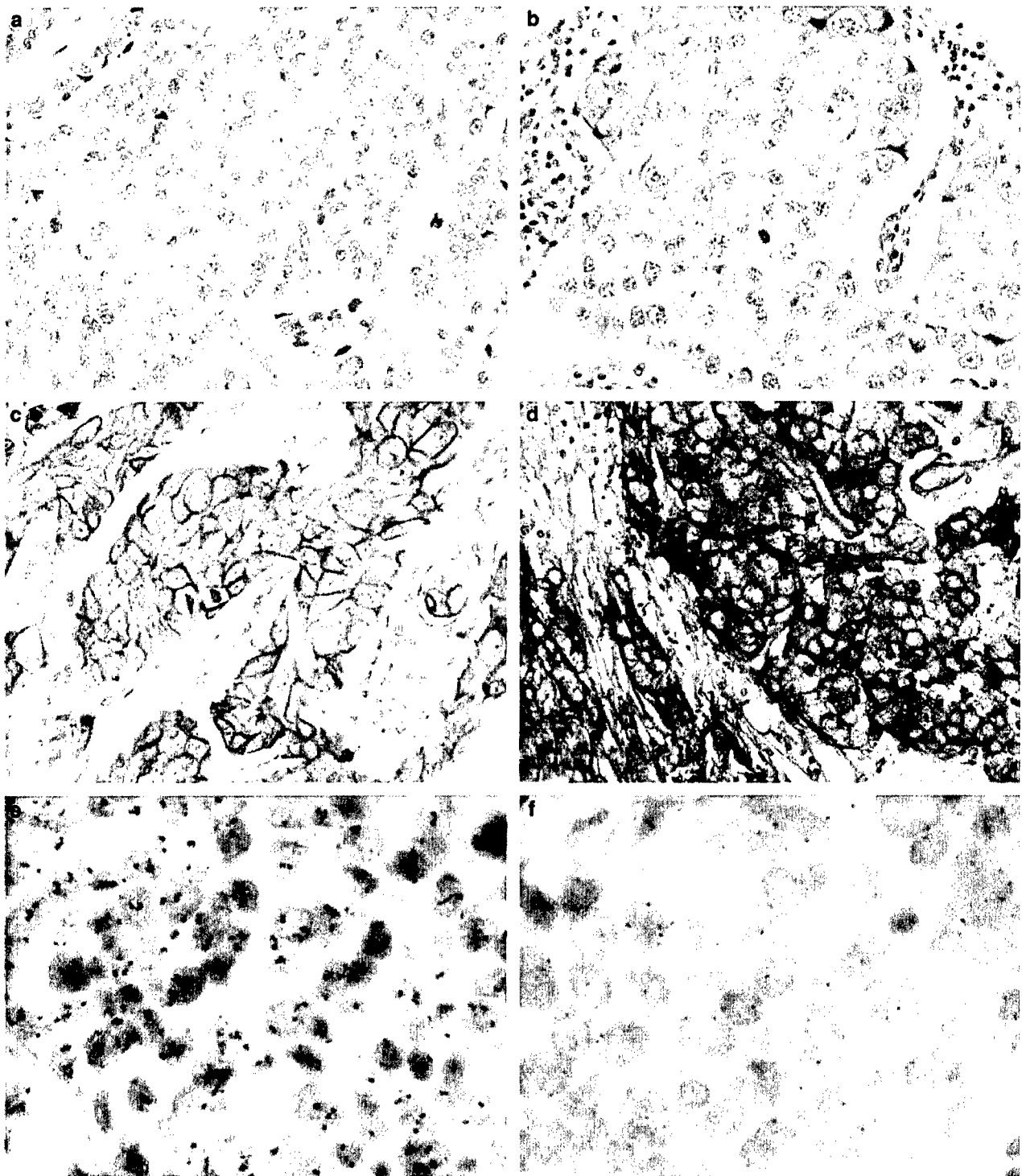


Figure 1 EGFR protein expression by immunohistochemistry and gene amplification by CISH. (a) 0 by immunohistochemistry, (b) 1 + by immunohistochemistry, (c) 2 + by immunohistochemistry, (d) 3 + by immunohistochemistry, (e) gene amplification (10–12 gene copies per nucleus) by CISH, (f) no gene amplification (2–3 gene copies per nucleus) by CISH.

Specific assays for the most frequent *EGFR* mutations in lung adenocarcinomas, exon 19 in-frame deletions and the exon 21 L858R point mutation, were used to analyze all *EGFR*-amplified tumors, and the one tumor with 3+ *EGFR* immu-

nohistochemistry without *EGFR* gene amplification. None of the tumors showed either of these hotspot mutations in the *EGFR* gene (Table 2).

We evaluated the clinical and pathologic features of *EGFR*-amplified breast cancers in an effort to

determine clinically relevant associations (Table 3). In all, 10 of these 11 tumors were poorly differentiated high-grade invasive ductal carcinoma, and one was a spindle cell metaplastic carcinoma with focal squamous differentiation. All of them were negative for estrogen receptor and progesterone receptor, but three of them were positive for HER-2 (Table 3). *EGFR* amplification appears to be inversely correlated with estrogen receptor expression. There was no correlation between *EGFR* amplification and *HER-2* amplification. Three of the 11 patients developed distant metastases at 40, 42, and 48 months, respectively, after the initial diagnoses (Table 3). The first two patients (No. 6 and 7) died of disease at 84 and 55 months, respectively, and the third patient (No. 9) is alive with lung and bone

metastases at 89 months. One other patient (No. 8) died of unrelated causes at 34 months. The mean follow-up of the 11 patients is 73 months. Owing to the limited number of informative cases, we were unable to determine whether *EGFR* amplification and/or *EGFR* overexpression is an independent prognostic indicator.

Discussion

Although the *EGFR* gene was identified more than two decades ago,²¹ clinical interest in the gene has recently been heightened by the discovery of *EGFR* inhibitors. In 1996, Yang *et al*²² demonstrated that treatment with genistein, an inhibitor of tyrosine kinase activity, inhibited EGF-induced tyrosine phosphorylation and degradation of *EGFR* in HepG2 cells, suggesting that tyrosine kinase activity is required for either the internalization or the degradation of EGF-*EGFR* receptor complexes. The use of *EGFR* kinase inhibitors has recently received FDA approval for use in cancer therapy.

In this study, we used CISH to detect *EGFR* gene amplification in breast carcinomas. Our data revealed that *EGFR* gene amplification is an infrequent event in breast cancer, occurring in only 6% of tumors. This percentage is in the middle of the range reported by the few previous studies that have examined *EGFR* copy number in breast cancer (0.8–14%).^{23,24}

EGFR overexpression was seen in 6% tumors in our current study, which correlated well with gene amplification. Most studies that have reported a higher percentage of *EGFR* overexpression have not evaluated gene amplification.^{11–13} Differences in the prevalence of *EGFR* overexpression reported by different studies may be due to variations in techniques and type of antibodies used, criteria for determining overexpression and interobserver variability. For example, Harris *et al*¹¹ measured *EGFR* in 221 primary breast cancers by ligand

Table 2 Detailed data on *EGFR* protein expression by immunohistochemistry, mRNA level, gene copy number by CISH, and mutation status in tumors with *EGFR* amplification (*n* = 11)

Case no.	CISH ^a	Immunohistochemical scores	mRNA ^b	Hotspot mutations ^c
1	7	1+	NA	NF
2	7	2+	NA	NF
3	8	3+	NA	NF
4	10	3+	NA	NF
5	11	3+	NA	NF
6	12	3+	34	NF
7	15	2+	5.3	NF
8	15	3+	NA	NF
9	15	3+	<2	NF
10	15	3+	<2	NF
11	18	3+	41	NF

^aData represent *EGFR* gene copy number per nucleus.

^bData represent fold increase above average mRNA level of *EGFR*-nonamplified tumors derived from Affymetrix U133A chip hybridizations. Calculations of relative *EGFR* transcript levels were based on data from Affymetrix probe set 201984_s_at.

^cMutations in *EGFR* exon 19 (short in-frame deletions) and exon 21 (L858R mutation).

CISH: chromogenic *in situ* hybridization; NA: not available; NF: not found.

Table 3 Detailed clinical and pathologic data in tumors with *EGFR* amplification (*n* = 11)

Case no.	Age (years)	Stage	Tumor type	Architectural grade	Nuclear grade	HER-2 FISH ^a	HER-2 IHC	ER	PR	Recurrence (months)	Survival (months)
1	44	3C	Ductal	3	3	3.8	3+	—	—	None	38 (NED)
2	47	2B	Ductal	3	2	10.7	3+	—	—	None	141 (NED)
3	40	2B	Ductal	3	3	NA	0	—	—	None	74 (NED)
4	41	3C	Ductal	3	3	1.0	0	—	—	None	40 (NED)
5	50	2B	Ductal	3	3	NA	0	—	—	None	91 (NED)
6	58	2A	Ductal	3	2	NA	0	—	—	40	84 (DOD)
7	52	2B	Ductal	3	3	1.5	1+	—	—	42	55 (DOD)
8	92	2A	Ductal	3	3	5.4	3+	—	—	None	34 (DOC)
9	61	2B	Metaplastic	3		1.0	0	—	—	48	89 (AWD)
10	64	2A	Ductal	3	3	NA	0	—	—	None	92 (NED)
11	54	3A	Ductal	3	3	NA	1+	—	—	None	66 (NED)

^aData represent ratio of *HER-2*/chromosome 17 copy numbers.

IHC: immunohistochemistry; ER: estrogen receptor; PR: progesterone receptor; FISH: fluorescence *in situ* hybridization; NA: not available; NED: no evidence of disease; DOD: dead of disease; AWD: alive with disease; DOC: dead of other causes; —: negative.

binding with ^{125}I -labelled EGF, and high-affinity sites were quantitated. Tsutsui *et al*¹² used a primary EGFR monoclonal antibody (Kyokutou Seiyaku, Tokyo, Japan) for assessing EGFR expression, and interpreted overexpression as 'tumors exhibiting definite staining of the cancer cells'. In our current study, tumors with 1+ staining intensity were interpreted as negative for overexpression. Our stringent criteria in defining EGFR overexpression appeared to be the major contributing factor to the apparent low prevalence of EGFR overexpression among breast carcinomas in this study.

We found no correlation of *EGFR* amplification and HER-2 status. Of the 11 tumors showing *EGFR* gene amplification, three tumors (27%) showed HER-2 overexpression. These three tumors also showed *HER-2* gene amplification. This proportion of HER-2 positivity approximates the expected percentage in breast cancers in general. The 11 *EGFR*-amplified tumors were uniformly estrogen receptor/progesterone receptor-negative, consistent with findings by other investigators.²³

There are contradictory reports in the literature on the prognostic significance of EGFR overexpression and its relationship with known prognostic factors.^{25–28} In the only study that examined the survival impact of *EGFR* gene amplification, no correlation was found.²³ The clinical significance of *EGFR* amplification and/or EGFR overexpression could not be independently evaluated in our current study due to the small number of informative cases.

Low-level amplification of *EGFR* in concert with *EGFR* mutation is present in some lung adenocarcinoma cell lines²⁹ and we (M Ladanyi, unpublished data) and others have also observed that many clinical lung cancer samples show evidence of copy number gains of the mutant allele.³⁰ Based on these considerations, it was of interest to screen the *EGFR*-amplified tumors in the present study for the activating mutations in exon 19 and 21 that are commonly detected in lung cancers. However, no mutations were found.

EGFR gene amplification generally results in increased protein expression in breast carcinomas. Apparent EGFR protein overexpression without gene amplification occurred in only 2% of tumors in this study, and its mechanism needs to be further investigated. Overall, approximately 6% of breast carcinomas show moderate- to low-level *EGFR* amplification associated with genuine EGFR protein overexpression. A small minority of breast cancers could be responsive to EGFR-targeted therapy, and this carefully selected subset of patients should be considered for clinical trials evaluating EGFR antibodies or small inhibitory molecules.

References

- 1 Cohen S, Ushiro H, Stoscheck C, *et al*. A native 170,000 epidermal growth factor receptor-kinase complex from

- shed plasma membrane vesicles. *J Biol Chem* 1982; 257:1523–1531.
- 2 McCune BK, Earp HS. The epidermal growth factor receptor tyrosine kinase in liver epithelial cells. The effect of ligand-dependent changes in cellular location. *J Biol Chem* 1989;264:15501–15507.
- 3 Alroy I, Yarden Y. The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions. *FEBS Lett* 1997;410:83–86.
- 4 Burgering BM, Coffey PJ. Protein kinase B (c-Akt) in phosphatidylinositol-3-OH kinase signal transduction. *Nature* 1995;376:599–602.
- 5 Liu W, Li J, Roth RA. Heregulin regulation of Akt/protein kinase B in breast cancer cells. *Biochem Biophys Res Commun* 1999;261:897–903.
- 6 Muthuswamy SK, Gilman M, Brugge JS. Controlled dimerization of ErbB receptors provides evidence for differential signaling by homo- and heterodimers. *Mol Cell Biol* 1999;19:6845–6857.
- 7 Chan TO, Rittenhouse SE, Tsichlis PN. AKT/PKB and other D3 phosphoinositide-regulated kinases: kinase activation by phosphoinositide-dependent phosphorylation. *Annu Rev Biochem* 1999;68:965–1014.
- 8 Lynch TJ, Bell DW, Sordella R, *et al*. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 2004;350:2129–2139.
- 9 Paez JG, Janne PA, Lee JC, *et al*. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 2004;304:1497–1500.
- 10 Pao W, Miller V, Zakowski M, *et al*. EGF receptor gene mutations are common in lung cancers from 'never smokers' and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc Natl Acad Sci USA* 2004;101:13306–13311.
- 11 Harris AL, Nicholson S, Sainsbury JR, *et al*. Epidermal growth factor receptors in breast cancer: association with early relapse and death, poor response to hormones and interactions with neu. *J Steroid Biochem* 1989;34:123–131.
- 12 Tsutsui S, Kataoka A, Ohno S, *et al*. Prognostic and predictive value of epidermal growth factor receptor in recurrent breast cancer. *Clin Cancer Res* 2002;8:3454–3460.
- 13 Walker RA, Dearing SJ. Expression of epidermal growth factor receptor mRNA and protein in primary breast carcinomas. *Breast Cancer Res Treat* 1999;53: 167–176.
- 14 Kallioniemi OP, Wagner U, Kononen J, *et al*. Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum Mol Genet* 2001;10:657–662.
- 15 Kononen J, Bubendorf L, Kallioniemi A, *et al*. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844–847.
- 16 Skacel M, Skilton B, Pettay JD, *et al*. Tissue microarrays: a powerful tool for high-throughput analysis of clinical specimens: a review of the method with validation data. *Appl Immunohistochem Mol Morphol* 2002;10:1–6.
- 17 Bhargava R, Lal P, Chen B. Chromogenic *in situ* hybridization for the detection of HER-2/neu gene amplification in breast cancer with an emphasis on tumors with borderline or low-level amplification. Does it measure up to fluorescent *in situ* hybridization? *Am J Clin Pathol* 2005;123:237–243.

- 18 Bhargava R, Lal P, Chen B. Feasibility of using tissue microarrays for the assessment of HER-2 gene amplification by fluorescence *in situ* hybridization in breast carcinoma. *Diagn Mol Pathol* 2004;13:213–216.
- 19 LaTulippe E, Satagopan J, Smith A, *et al*. Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* 2002;62:4499–4506.
- 20 Pan Q, Pao W, Ladanyi M. Rapid PCR-based detection of epidermal growth factor receptor gene mutations in lung adenocarcinomas. *J Mol Diagn* (in press).
- 21 Shimizu N, Behzadian MA, Shimizu Y. Genetics of cell surface receptors for bioactive polypeptides: binding of epidermal growth factor is associated with the presence of human chromosome 7 in human-mouse cell hybrids. *Proc Natl Acad Sci USA* 1980;77:3600–3604.
- 22 Yang EB, Wang DF, Mack P, *et al*. Genistein, a tyrosine kinase inhibitor, reduces EGF-induced EGF receptor internalization and degradation in human hepatoma HepG2 cells. *Biochem Biophys Res Commun* 1996;224:309–317.
- 23 Al Kuraya K, Schraml P, Torhorst J, *et al*. Prognostic relevance of gene amplifications and coamplifications in breast cancer. *Cancer Res* 2004;64:8534–8540.
- 24 Ro J, North SM, Gallick GE, *et al*. Amplified and overexpressed epidermal growth factor receptor gene in uncultured primary human breast carcinoma. *Cancer Res* 1988;48:161–164.
- 25 Cerra M, Cecco L, Montella M, *et al*. Epidermal growth factor receptor in human breast cancer comparison with steroid receptors and other prognostic factors. *Int J Biol Markers* 1995;10:136–142.
- 26 Fox SB, Smith K, Hollyer J, *et al*. The epidermal growth factor receptor as a prognostic marker: results of 370 patients and review of 3009 patients. *Breast Cancer Res Treat* 1994;29:41–49.
- 27 Pirinen R, Lipponen P, Syrjanen K. Expression of epidermal growth factor receptor (EGFR) in breast cancer as related to clinical, prognostic and cytometric factors. *Anticancer Res* 1995;15:2835–2840.
- 28 Toi M, Nakamura T, Mukaida H, *et al*. Relationship between epidermal growth factor receptor status and various prognostic factors in human breast cancer. *Cancer* 1990;65:1980–1984.
- 29 Amann J, Kalyankrishna S, Massion PP, *et al*. Aberrant epidermal growth factor receptor signaling and enhanced sensitivity to EGFR inhibitors in lung cancer. *Cancer Res* 2005;65:226–235.
- 30 Kosaka T, Yatabe Y, Endoh H, *et al*. Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications. *Cancer Res* 2004;64:8919–8923.



Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors

Andy J. Minn,^{1,2} Yibin Kang,¹ Inna Serganova,³ Gaorav P. Gupta,¹ Dilip D. Giri,⁴ Mikhail Doubrovin,³ Vladimir Ponomarev,³ William L. Gerald,⁴ Ronald Blasberg,³ and Joan Massagué^{1,5}

¹Cancer Biology and Genetics Program, ²Department of Radiation Oncology, ³Department of Neurology, ⁴Department of Pathology, and ⁵Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, New York, New York, USA.

We used bioluminescence imaging to reveal patterns of metastasis formation by human breast cancer cells in immunodeficient mice. Individual cells from a population established in culture from the pleural effusion of a breast cancer patient showed distinct patterns of organ-specific metastasis. Single-cell progenies derived from this population exhibited markedly different abilities to metastasize to the bone, lung, or adrenal medulla, which suggests that metastases to different organs have different requirements. Transcriptomic profiling revealed that these different single-cell progenies similarly express a previously described “poor-prognosis” gene expression signature. Unsupervised classification using the transcriptomic data set supported the hypothesis that organ-specific metastasis by breast cancer cells is controlled by metastasis-specific genes that are separate from a general poor-prognosis gene expression signature. Furthermore, by using a gene expression signature associated with the ability of these cells to metastasize to bone, we were able to distinguish primary breast carcinomas that preferentially metastasized to bone from those that preferentially metastasized elsewhere. These results suggest that the bone-specific metastatic phenotypes and gene expression signature identified in a mouse model may be clinically relevant.

Introduction

Cancer metastases are responsible for the majority of cancer-related deaths. A widely held hypothesis is that cancer metastasis arises from rare cells in the primary tumor that acquire the ability to progress through sequential steps necessary to grow at a distant site (1, 2). Some of these sequential steps include invasion through extracellular matrix, intravasation, survival in the circulation, extravasation into a distant site, and progressive growth at that site. Consistent with the multistep nature, there is experimental and clinical evidence to suggest that metastasis is an inefficient process whereby the vast majority of circulating tumor cells are not able to progressively grow at distant sites (3–6). Related to this is the observation that metastatic cells exhibit tissue tropism, preferring to grow in certain organs in a way that cannot be explained by circulatory patterns alone. In breast cancer, for example, metastasis affects the bone and the lung, and less frequently the liver, brain, and adrenal medulla. Although the genetic basis of these metastatic properties is poorly understood, acquisition of the ability to complete each step involved in metastasis is thought to be driven by the accumulation of genetic mutations that may result in a rare cell's acquisition of a full complement of these mutations relatively late during the evolution of the primary tumor (1).

Recently, the development of DNA microarray technology, which allows for genome-wide transcriptomic profiling, has provided new insight into the genetic basis of metastasis. Studies using primary

tumor material have identified a gene expression signature for breast cancer metastasis consisting of a set of 70 genes (7, 8). The presence of this “poor-prognosis” signature in the primary tumor from early stage breast cancer patients is highly prognostic for the development of distant metastasis and overall survival. Work using adenocarcinoma metastases and unmatched primary tumors from breast and other tumor types has revealed similar findings (9).

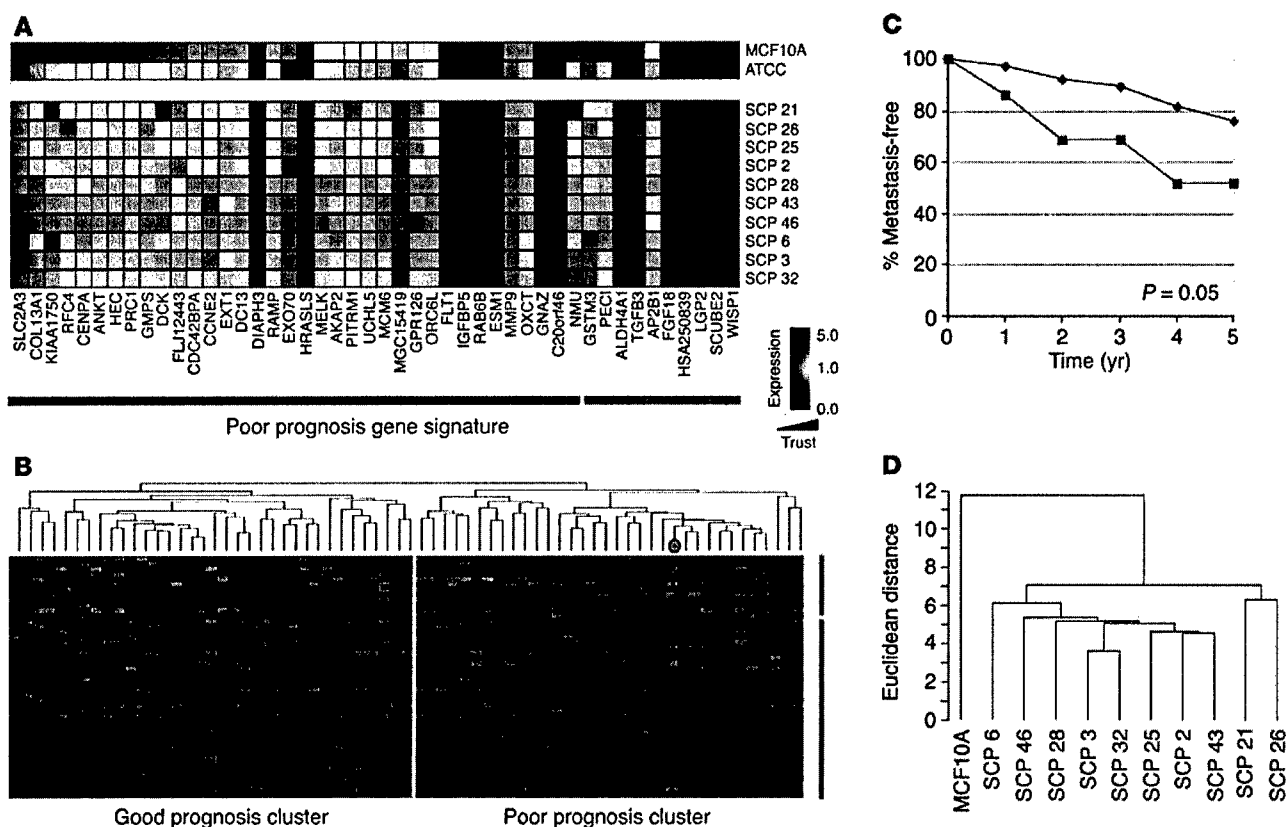
The fact that the poor-prognosis signature from early-stage primary cancers can be used to predict the development of distant metastasis has been interpreted as challenging the traditional model of metastasis because it suggests that metastatic cells may result from many of the early oncogenic events that drive primary tumor growth rather than developing from late-arising, rare cells that accumulate genomic alterations specific for metastasis (10). Other researchers have maintained the existence of distinct metastasis genes and have argued that a poor-prognosis signature may result from the aggregate contribution of these genes by subpopulations of cells that aberrantly express some but not all of the multiple genes required to complete metastasis (11). Thus, the cell that contains the full complement of metastasis-enabling genes still may be rare. Regardless, the ability of the poor-prognosis genes to directly mediate metastasis remains unknown.

Using *in vivo* selection of organ-specific metastatic cells from the human breast cancer cell line MDA-MB-231, we recently identified and functionally validated a set of genes that specifically mediate osteolytic bone metastasis in the mouse (12). Cells that express these genes and that are capable of bone metastasis pre-exist within the MDA-MB-231 parent line, which as a population already carries the poor-prognosis signature. This cell line was originally established as the total outgrowth of cells derived from a pleural effusion of a patient who relapsed years after removal of the primary tumor (13). In the present study, we investigate the

Nonstandard abbreviations used: ATCC, American Type Culture Collection; BLI, bioluminescence imaging; eGFP, enhanced green fluorescent protein; HSV1-TK, herpes simplex virus 1 thymidine kinase; R index, reproducibility index; ROI, region of interest; SCP, single cell-derived progeny; TGL, thymidine kinase; GFP, luciferase.

Conflict of interest: The authors have declared that no conflict of interest exists.

Citation for this article: *J. Clin. Invest.* 115:44–55 (2005).
doi:10.1172/JCI200522320.

**Figure 1**

SCPs from MDA-MB-231 cells have a poor-prognosis gene expression signature. **(A)** Microarray expression data of 46 of the 70 poor-prognosis genes (7) that are present on the Affymetrix U133A GeneChip for the MCF10A normal breast epithelial cell line, parental MDA-MB-231 cell line, and various SCPs from MDA-MB-231. Each column represents a gene (denoted along the bottom) and each row represents a cell line (denoted along the right). Genes of the poor-prognosis signature that are expressed at higher levels in poor-prognosis tumors are above the red line, and those that are underexpressed are above the green line. Genes with low trust values due to low or absent expression are shaded in darker colors (Trust; wedge). **(B)** Microarray expression data of primary human breast carcinoma from 63 patients treated at our institution who had at least 5 years of clinical follow-up and/or developed metastatic disease. Hierarchical clustering of the patients' data was performed with the 46 poor-prognosis genes. Each column represents a patient and each row, a gene. The MDA-MB-231 cell line was included and is denoted by a blue dot in the dendrogram. Those patients in the good-prognosis versus the poor-prognosis cluster are separated by the yellow line. **(C)** Five-year metastasis-free survival data for the 63 patients classified according to the hierarchical clustering described in **B**. The P value shown in the graph was calculated by the χ^2 test. **(D)** Dendrogram showing hierarchical clustering of the SCPs and MCF10A using the poor-prognosis genes. A scale of the distance metric used is shown on the left.

relationship between this bone metastasis signature, the general poor-prognosis signature, and the metastatic activity of individual cells from the parental population and of a cohort of metastatic human primary tumors.

Results

Similar poor-prognosis gene expression signatures in different single cell-derived progenies. The poor-prognosis gene expression signature for breast cancer, which can be used to predict the development of distant metastasis, consists of 70 genes, 58 of which are upregulated and 18 of which are downregulated, and correlates closely with negative estrogen receptor status (7). Most tumors in the poor-prognosis group have only a fraction (on average, approximately one third) of the 70 gene expression events that constitute the poor-prognosis signature. Furthermore, these gene expression events often show extensive variation among different tumors with a poor prognosis. We recently reported that MDA-MB-231 cells, as directly obtained from the American Type Culture Collection (ATCC), also have the

poor-prognosis signature. Of the 70 genes from this signature, 46 were present on the Affymetrix U133A GeneChip that we used for our microarray analysis (Figure 1A). Of the 58 upregulated genes of the poor-prognosis signature, 36 were present on this microarray. Compared with the MCF10A cell line derived from nonmalignant human breast epithelium, the majority of these 36 genes were upregulated in parental MDA-MB-231 cells. Of the 18 downregulated genes from the poor-prognosis signature, 10 were present on the U133A GeneChip. Consistent with downregulation in poor-prognosis tumors, 7 of the 10 had low trust values due to their low or absent expression.

To further confirm that MDA-MB-231 cells have a poor-prognosis gene expression signature, we compared the transcriptomic profile of these cells with that of a cohort of primary breast carcinomas from patients treated at the Memorial Sloan-Kettering Cancer Center. All of these patients had at least 5 years of clinical follow-up or had developed metastatic disease. Hierarchical clustering using the poor-prognosis gene expression signature (7) separated these tumors

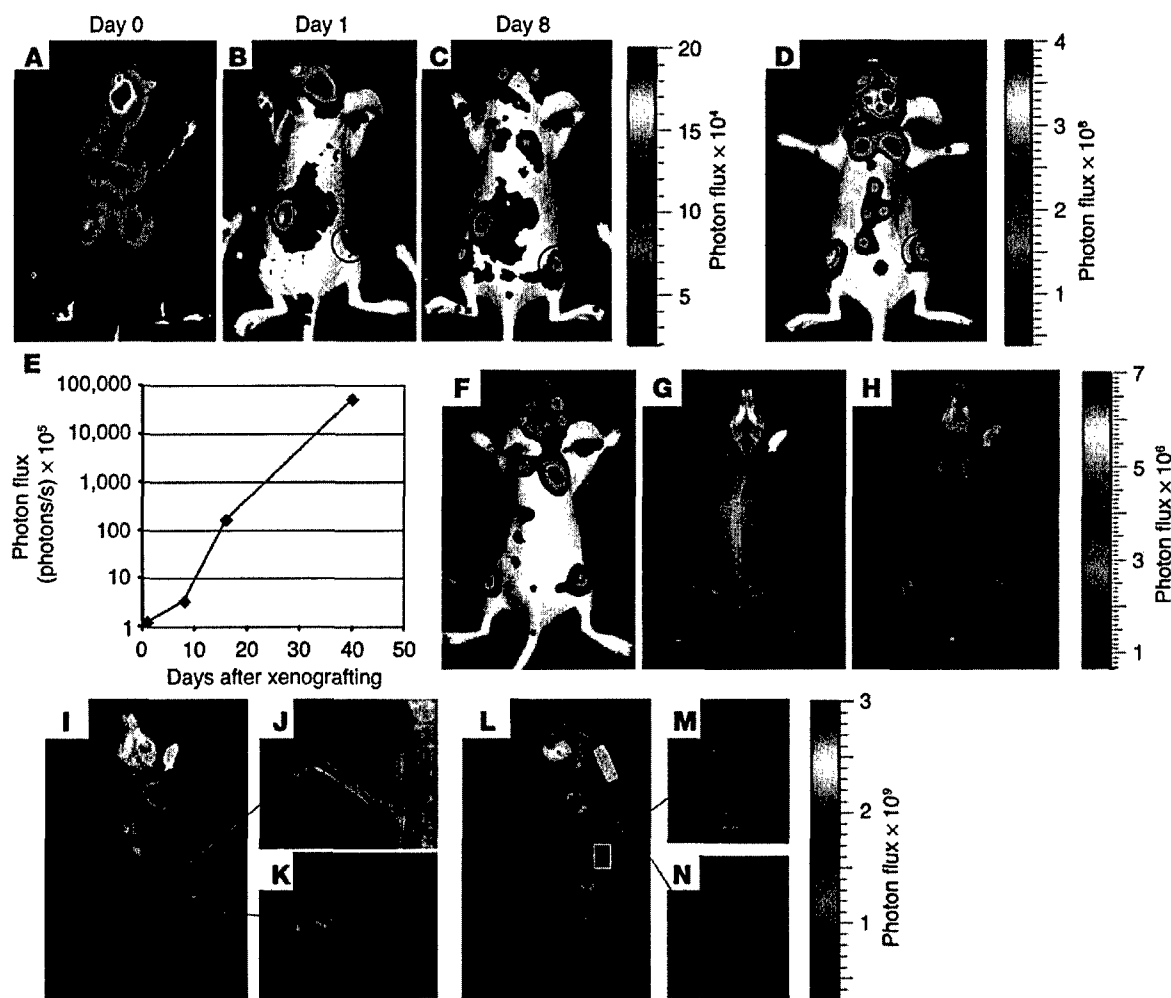


Figure 2

Noninvasive BLI to monitor the development of osteolytic metastases from the same mouse. (A–D) SCP2, a highly metastatic clone from MDA-MB-231, was transduced with the luciferase-containing TGL reporter gene and was injected into the left cardiac ventricle of an immunodeficient mouse. At the indicated times after xenografting, the bioluminescence signal was captured. The intensity of the signal, measured as photon flux, is shown as a color scale. Images for days 0, 1, and 8 are displayed on the same scale, while the day-35 image is shown on a different scale due to the exponential growth of the metastases. A metastasis to the right hindlimb is circled in red. (E) The growth kinetics of the right hindlimb metastasis outlined by the red circle shown in B–D was quantified by measurement of photon flux. (F–H) A bioluminescence image (F) and a skeletal x-ray image (G) were obtained on day 16 after xenografting. Images were superimposed (H) to demonstrate registration of the bioluminescence signals with skeletal anatomy. (I–N) A superimposed image from day 45 (I and L) reveals extensive areas of osteolytic destruction that correspond to bioluminescence signals. Magnification of regions outlined in red shows involvement of the femur/tibia, iliac crest of the pelvis, and the sacrum (J and K), in addition to the vertebrae (M and N). The bioluminescence signal from the region outlined in yellow on the left lateral projection (L) does not overlap with skeletal structures and originates from the adrenal gland (Figure 3, J–M).

into two major clusters, one cluster corresponding to patients with a poor-prognosis signature and the other representing those with a “good-prognosis” signature (Figure 1B). Consistent with previous reports, patients in our cohort with a poor-prognosis signature had a significantly worse 5-year metastasis-free survival than those with the good-prognosis signature (Figure 1C). MDA-MB-231 cells fall squarely within this poor-prognosis group (Figure 1B). Thus, MDA-MB-231 cells express a typical poor-prognosis tumor profile.

Among the questions raised by these observations is whether the particular set of poor-prognosis gene expression events presented by a poor-prognosis tumor reflects the presence of this particular pattern in the majority of malignant cells of the tumor or if it reflects contributions from different cells in the population.

To address this question in the MDA-MB-231 case, we used various single cell-derived progenies (SCPs) obtained from single-cell cloning and analyzed them for the presence of a poor-prognosis signature. Although there was some variation among the SCPs in the expression levels of the genes that comprised the signature, the SCPs maintained a set of poor-prognosis gene expression events similar to that found in the ATCC population from which they were derived (Figure 1A). A dendrogram of the SCPs using the poor-prognosis gene set confirmed that the distance metric between the SCPs was significantly less than the distance metric between the whole group of SCPs and MCF10A (Figure 1D).

Flow cytometry analysis of the parental MDA-MB-231 cell population indicated that approximately 10% of cells in this population

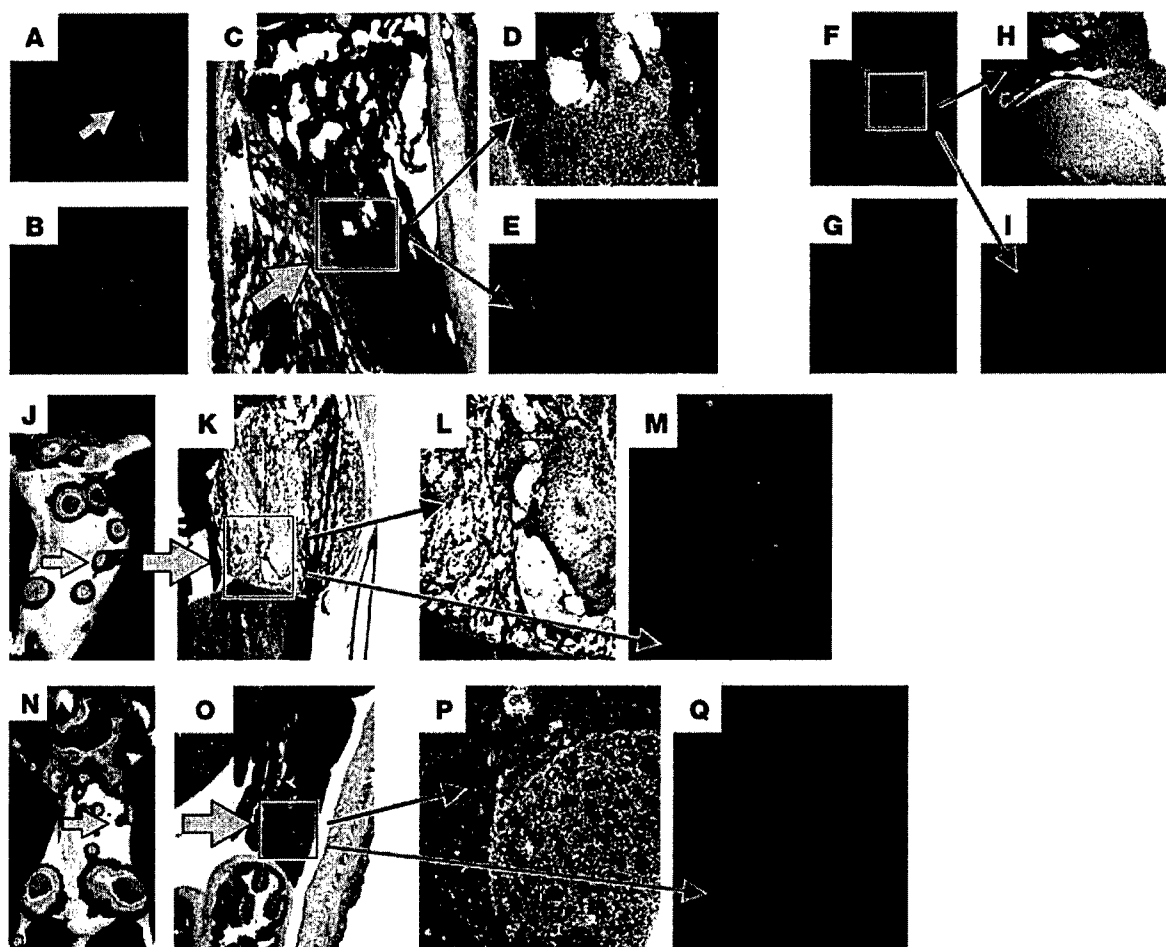


Figure 3

Verification of macroscopic and microscopic metastases by fluorescence histology. (A–I) A pathological fracture involving the proximal tibia (A–E) or vertebrae (F–I) is demonstrated by skeletal x-ray (A and B) and an overlay of this x-ray with BLI (B and G) from the same mouse as that described in Figure 2. To confirm metastases, we performed whole-mount frozen sectioning. Regions corresponding to the fractured tibia and vertebra were analyzed by H&E staining (C, D, and H) or unstained sections were analyzed for GFP fluorescence (E and I). (J–M) A lateral projection of a bioluminescence image from day 45 (J) corresponding to the same image as that in Figure 2L reveals a signal originating from the adrenal gland (green arrow), as shown by H&E staining (K). Magnification of the boxed region in K (L) and GFP fluorescence (M) of the left adrenal gland are shown. (N–Q) Inspection of organs in the left upper abdominal quadrant with areas of bioluminescence signal (N) reveals a focus of tumor growth in the pancreas (O). Magnification of the boxed region in O (P) and GFP fluorescence (Q) are shown.

expressed CXCR4 (data not shown), a product representative of the bone metastasis gene expression signature (12). A similar percentage of SCPs were found to overexpress CXCR4 (12). Thus, based on these criteria at least, our single-cell cloning process did not introduce bias in the selection of cell clones representing the parental population.

Noninvasive bioluminescence imaging of metastases. After intracardiac injection of parental MDA-MB-231 cells into immune-deficient mice, approximately 30% will develop osteolytic bone metastasis that is evident by skeletal x-ray imaging (12). Subpopulations that are more osteolytic than the parental population have been obtained through a process of *in vivo* selection for bone metastasis or by isolation of SCPs from parental MDA-MB-231 cells. However, the sensitivity of skeletal x-ray in detecting nonosseous metastasis is poor. Likewise, findings at necropsy may also fail to reveal small and/or anatomically inconspicuous lesions. Indeed, at necropsy, MDA-MB-231 cells are infrequently found to have metastasized to nonosseous organs such as the adrenal medulla.

In order to better characterize the overall metastatic properties of MDA-MB-231 SCPs and their relationships to both the poor prognosis and the bone metastasis gene sets, we used luciferase-based, noninvasive bioluminescence imaging (BLI) and fluorescence microscopy using a novel triple-modality reporter gene, thymidine kinase, GFP, luciferase (TGL) (14). This artificial gene encodes a triple fusion protein with herpes simplex virus 1 thymidine kinase (HSV1-TK) fused to the N terminus of enhanced green fluorescent protein (eGFP) and firefly luciferase fused to the C terminus of eGFP. When transduced into cells, HSV1-TK allows for nuclear imaging, eGFP can be utilized for fluorescence, and luciferase allows for BLI.

SCP2 is a single cell-derived population of MDA-MB-231 cells that produces aggressive osteolytic lesions by 8 weeks after left ventricular cardiac injection into immunodeficient mice. As a test of the sensitivity and resolution of the TGL reporter gene, we transduced SCP2 with the TGL reporter and monitored the development

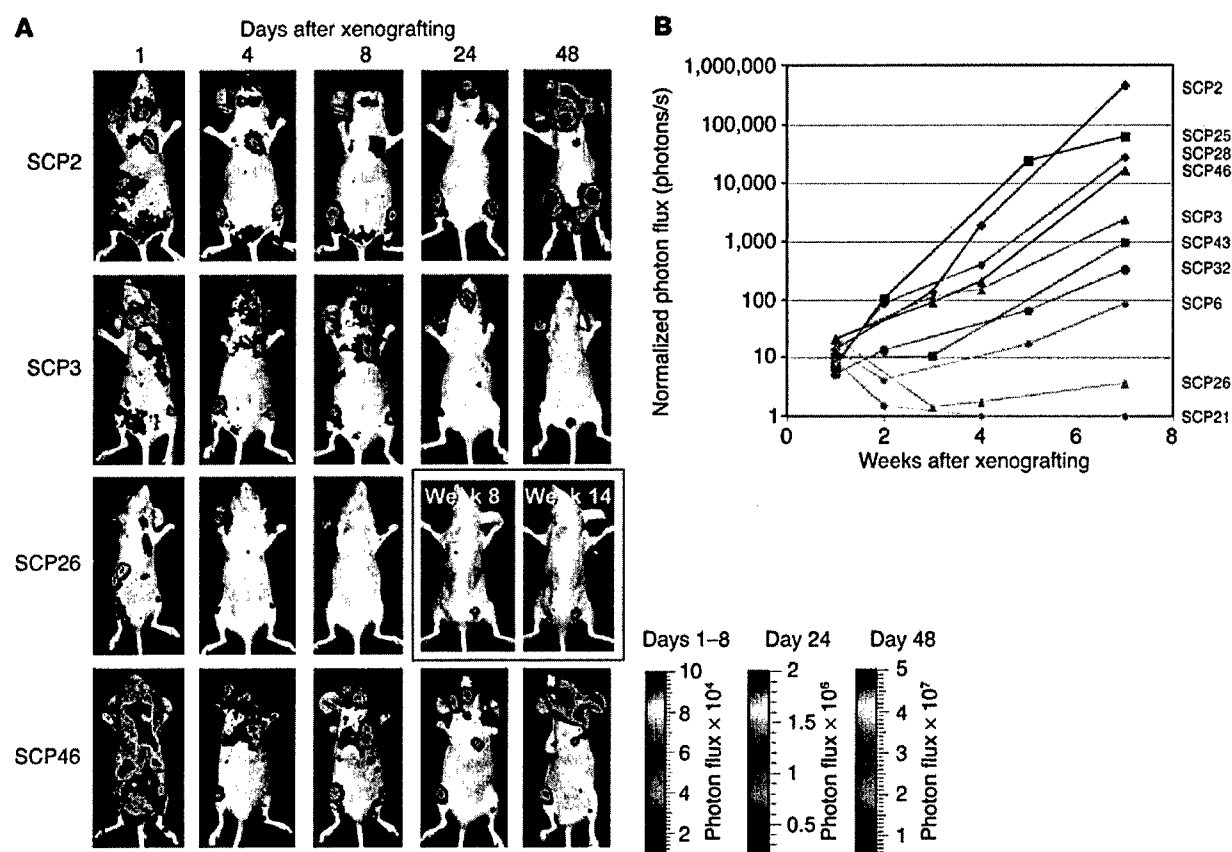


Figure 4

SCPs exhibit different abilities to metastasize to bone. (A and B) Each of the SCPs was labeled with the TGL reporter, and 1×10^5 cells were injected into the left cardiac ventricle. At the indicated days after xenografting, bioluminescence images were acquired. (A) Representative mice injected with a representative set of SCPs are shown in the supine position. The intensity of the signal from days 1, 4, and 8 are on equivalent scales, while day 24 and day 48 are each on separate scales due to increasing signal strength and to avoid signal saturation. (B) The normalized photon flux from the dominant signal originating from the hindlimbs, forelimbs, or pelvis of all the SCPs studied was measured over the indicated time course. SCPs were ranked according to their growth kinetics in either bone or lung. SCPs with a higher rank order for bone are shown in red, and those with a higher rank order for lung are shown in green. The bottom three SCPs for both bone and lung are classified as being the least metastatic and are shown in blue.

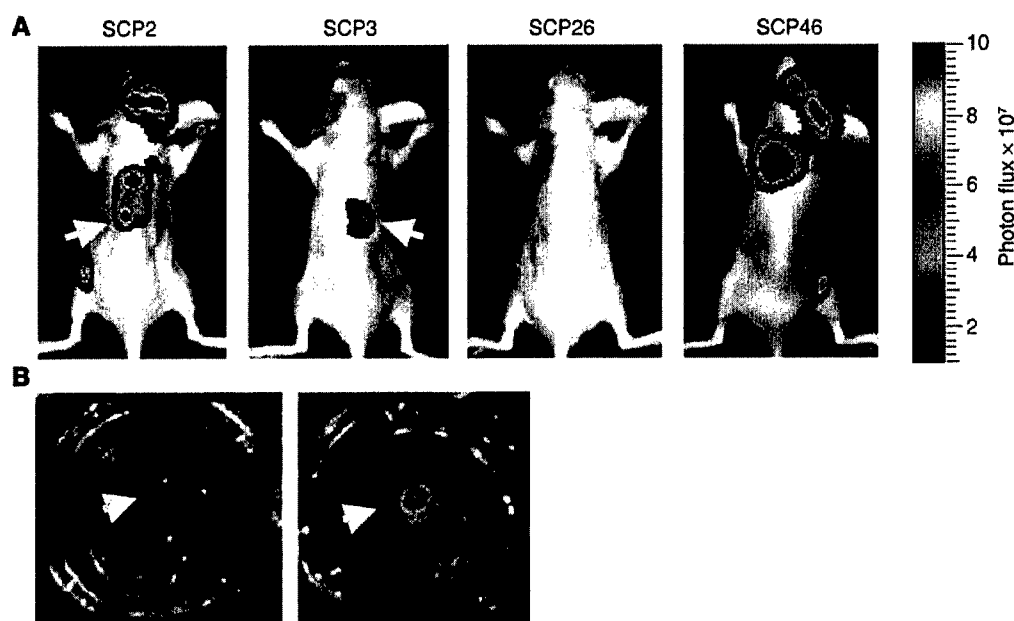
of osteolytic metastases. Shortly after the injection of 1×10^5 cells into the left cardiac ventricle, a diffuse whole-body bioluminescence signal was detected (Figure 2A). This signal followed systemic blood flow patterns, with areas of strongest signal probably corresponding to organs receiving the highest percentage of cardiac output; namely, kidney, liver, and brain. At day 1 after injection, much of the diffuse signal disappeared; however, foci of arrested tumor cells could be seen. These foci increased in number and intensity through the first week (Figure 2, A–C). In particular, an increasing signal could be detected in the hindlimbs that corresponded to primary areas for the development of osteolytic metastasis (Figure 2, B–D, red circles). This major hindlimb signal was quantified by measurement of photon flux and demonstrated logarithmic growth (Figure 2E).

Because bioluminescence signals could be correlated only with surface anatomy, we sought a way to assign major areas of bioluminescence to anatomical structures. At day 16 after injection, we overlaid the bioluminescence signal with skeletal x-ray images in order to analyze the correlation between areas of signal with skeletal anatomy. The majority of the signal overlapped well with bony structures, including the distal femur/proximal tibia, bony pelvis, scapula, vertebra, distal ulna, and skull (Figure 2, F–H). Although inspection of the

x-ray images at day 16 did not reveal evidence of osteolytic destruction at the sites of overlap, skeletal x-ray imaging of the same animal at day 45 demonstrated overlapping areas with extensive osteolytic destruction involving the distal femur/proximal tibia, iliac crest, sacrum, and vertebral body (Figure 2, I–N). Thus, these data suggest that BLI can be significantly more sensitive in detecting bone metastasis than x-ray imaging, as it allows monitoring of the development of bone metastasis from initial arrest to osseous destruction.

Verification of BLI by fluorescence histology. In order to examine the regions of osteolytic metastasis histologically and to search for other, less obvious sites of occult metastases, we used whole-mount frozen sectioning to look for tumor-derived GFP fluorescence by microscopy. Skeletal x-ray and BLI identified a pathological fracture of the tibia (Figure 3, A and B). H&E staining of sections corresponding to this region revealed tumor cells eroding through the cortex of the tibia (Figure 3, C and D), and GFP fluorescence of a serial section confirmed the metastasis (Figure 3E). Similarly, a collapsed vertebral body was also demonstrated to be due to growth of tumor cells through the bone and into the spinal canal (Figure 3, F–I).

Not all areas of bioluminescence signal could be overlaid with skeletal structures. For example, as shown on day 35 after xenografting,

**Figure 5**

Differential ability among SCPs to metastasize to the adrenal gland. (A) After intracardiac injection of individual SCPs, bioluminescence images were acquired and analyzed for signals originating from regions consistent with adrenal metastasis (arrows). Shown are representative mice at 7 weeks after injection with SCPs that show varying abilities to give rise to adrenal metastasis. (B) At necropsy, left and right adrenal glands (with the kidneys) were removed and were imaged ex vivo for bioluminescence. Arrows show the locations of the left and right adrenal glands, respectively, from a representative mouse with adrenal metastasis.

bioluminescence signals on both sides lateral to the vertebral column could be detected (Figure 2D). On a lateral projection, these signals lay anterior to the vertebrae (Figures 2L and 3J). At necropsy, enlarged and necrotic adrenal glands were noted (Figure 3, K and L), and fluorescence microscopy confirmed that this was due to metastasis (Figure 3M). In addition, careful analysis of whole-mount frozen sections also identified other nonosseous sites of microscopic metastases corresponding to weak regions of bioluminescence signal. For example, small foci of signal were noted in the upper left quadrant of the abdomen (Figure 3N). This signal was confirmed to be due to microscopic metastasis involving the pancreas (Figure 3, O–Q).

In total, these data demonstrate that the TGL reporter gene enables the use of a noninvasive method for tracking metastases from the initial arrest in distant organs to the development of gross lesions. The growth of these lesions can be quantified by measuring photon flux and confirmed by fluorescence microscopy. The sensitivity of the system is exemplified by the ability to detect and confirm microscopic metastases that would otherwise be overlooked by routine necropsy.

Differential bone-metastatic activity with a similar poor-prognosis signature. Empowered by the sensitivity of the TGL reporter system, we sought to fully characterize the metastatic phenotypes of the SCPs. To assess the metastatic activity that develops after hematogenous spread, we introduced each of the SCPs into the arterial circulation of immunodeficient mice by injection into the left cardiac ventricle. The major site of colonization and growth among the SCPs is the bone (hindlimbs, ribs, pelvis/sacrum, and skull/mandible) (Figure 4A). However, the SCPs displayed significant variation in their ability to grow in bone, even though the various SCPs proliferated in culture at comparable rates (data not shown). The dominant signals on the supine projections came from the hindlimbs and the

bones of the skull. For presentation purposes, the bioluminescence data from days 1–8 are displayed on the same scale and day 24 and day 48 are each displayed on a different scale. Comparisons within these groups across SCPs demonstrated that SCP2 and SCP46 were more metastatic to bone than are SCP3 and SCP26.

The dominant hindlimb lesion from the complete set of SCPs was quantified by measurement of photon flux, and the kinetics of growth are shown in Figure 4B. The aggressiveness of SCP2, SCP25, SCP28, and SCP46 in forming bone metastasis was shown by a 3- to 4-log growth of the dominant hindlimb lesion over the course of 7 weeks. Most of these mice became cachectic and were sacrificed. The aggressive nature of these SCPs is consistent with their expression of a previously described bone metastasis gene expression signature

Table 1
Adrenal metastases and SCPs

Progeny	Number of mice analyzed	Number of mice with adrenal metastases (%)
SCP2	4	2 (50)
SCP3	9	7 (78)
SCP25	4	1 (25)
SCP6	5	0
SCP32	4	0
SCP43	5	0
SCP21	5	0
SCP26	4	0
SCP28	5	0
SCP46	4	0

The presence of adrenal metastasis was determined for the entire cohort of SCPs.

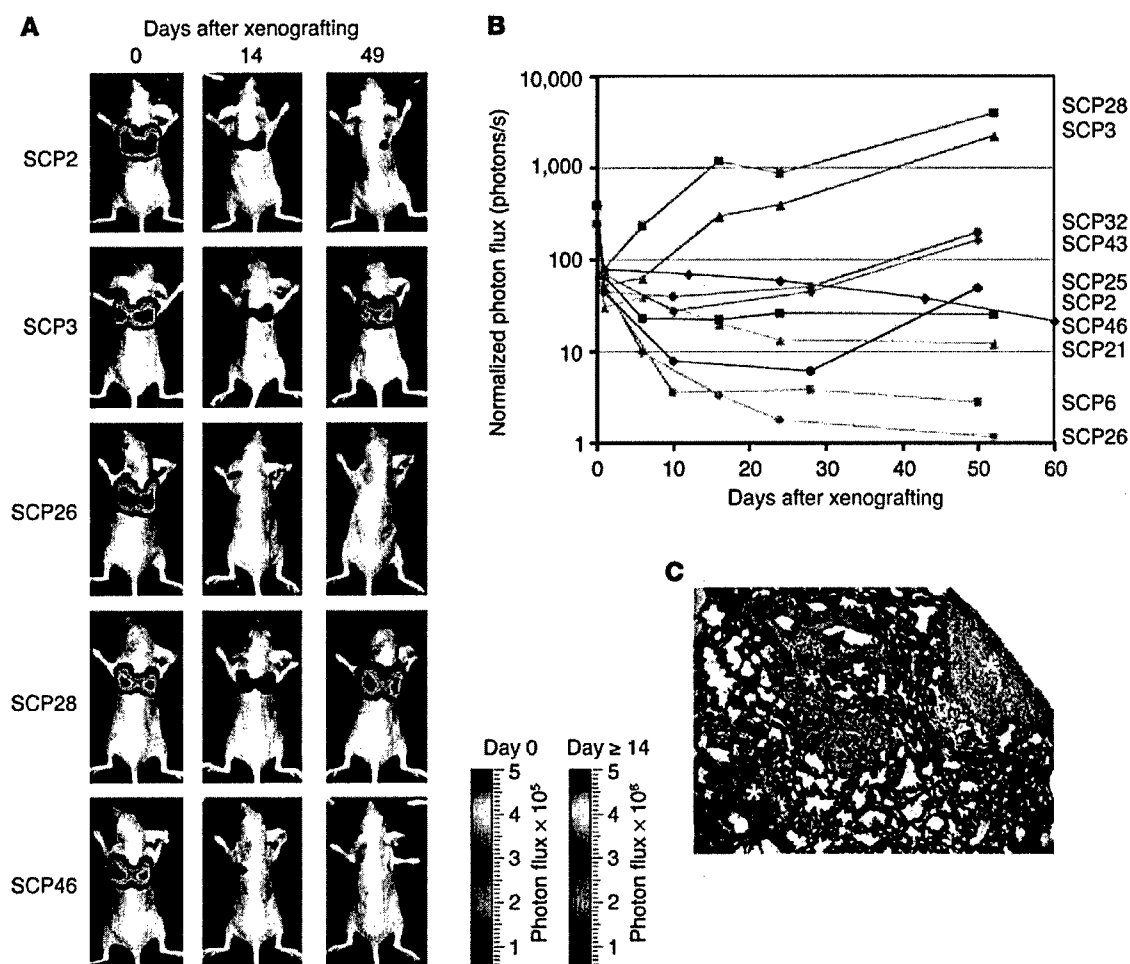


Figure 6

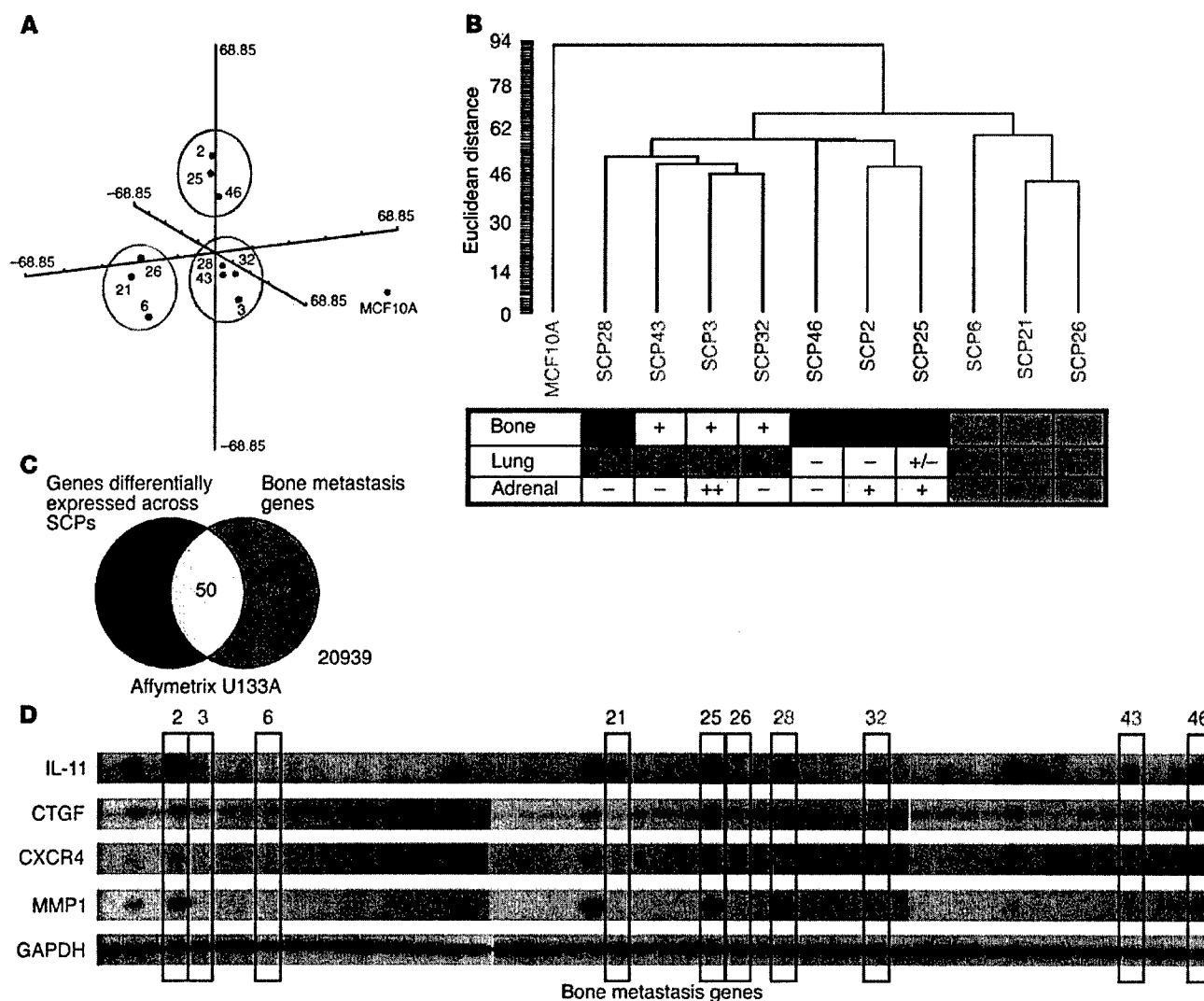
SCPs demonstrate different abilities to metastasize to the lung. (A–C) Each of the SCPs was labeled with the TGL reporter, and 2×10^5 cells were injected into the tail vein. At the indicated day after xenografting, bioluminescence images were acquired. (A) Representative mice injected with a representative set of SCPs are shown in the supine position. The intensity of the signal from day 0 is displayed on one scale, while that of days 14 and 49 (Day ≥ 14) are on a different scale due to increasing signal strength and to avoid signal saturation. (B) The normalized photon flux from the lung of all the SCPs studied was measured over the indicated time course. SCPs are color-coded as described in Figure 4B. (C) The lungs of SCPs that show growth in lung were analyzed histologically. A lung section from a representative SCP is shown stained for CD31, a marker for vascular endothelial cells, and counterstained with eosin. Asterisks mark regions of parenchymal tumor growth. The red arrow shows a CD31-positive blood vessel with an associated perivascular tumor growth pattern.

(12). However, SCP43, SCP3, and SCP32 were weaker in their metastatic growth to bone, while SCP6, SCP26, and SCP21 were the most weakly metastatic to bone. This reduction in bone metastasis ability correlated with the attenuation in expression of the bone metastasis genes (see Figure 7D). Interestingly, even among the weakest populations, we were able to detect the presence of bone metastasis. For example, at 14 weeks after xenografting of SCP26, a dormant metastatic focus within the hindlimbs was detectable in half of the mice (Figure 4A and Supplemental Figure 1; supplemental material available online with this article; doi:10.1172/JCI200522320DS1). Thus, these data demonstrate that the bone-metastatic activity of MDA-MB-231 cells does not correlate with the expression of their poor-prognosis signature but instead with the expression of our previously described bone metastasis gene set.

Different organ specificity of metastasis by different cells from the same population. After extensive analysis of metastatic growth by BLI,

whole-mount fluorescence microscopy, and micro-positron emission tomography (data not shown), we found bone to be the major site of tumor growth after arterial inoculation. In general, growth in other organs was rare, making comparable analysis unfeasible. However, one exception was metastatic growth in the adrenal gland, which occurred at an appreciable frequency. We were able to detect adrenal metastases in a minority of the SCPs by looking for dorsally located signals on either or both sides of the vertebral column that were suspicious for adrenal metastases (Figure 5A). These “suspicious” signals were confirmed at necropsy by gross inspection and/or ex vivo BLI (Figure 5B). Of the SCPs, SCP3 was the most consistent in producing adrenal metastasis (Table 1).

Due to size restrictions imposed by murine capillaries, human tumor cells are rarely able to pass from the arterial to the venous system (or vice versa) by way of the lungs (2). Therefore, we injected the SCPs into the tail vein in order to study the ability of SCPs to

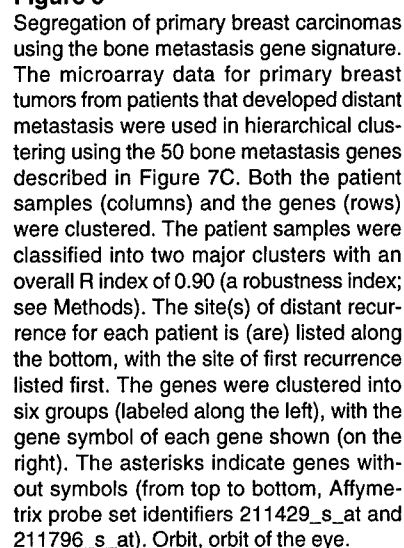
**Figure 7**

Genome-wide "unsupervised" classification of the SCPs correlates with metastatic phenotype. **(A)** A multidimensional scaling plot illustrates the relationship between the various SCPs and their primary metastatic tropism based on genes that are differentially expressed across the SCPs starting from the more than 22,000 present on the Affymetrix U133A GeneChip. SCPs are color-coded according to their primary metastatic tropism (green for lung, red for bone, and blue for weakly metastatic). The plot demonstrates that SCPs with the same primary metastatic tropism group together in 3-dimensional space. Each group is each enclosed in a circle. MCF10A is shown by itself (gold dot). **(B)** Hierarchical clustering of the SCPs based on genes differentially expressed reveals similar relationships and a similar association with metastatic tropism, as summarized in the table below the dendrogram. **(C)** A Venn diagram demonstrates the relationship between the genes differentially expressed across the SCPs and a previously described bone metastasis gene set. Of 1,267 differentially expressed genes, 50 of the 127 bone metastasis genes (102 are unique) overlap. **(D)** A Northern blot showing the expression levels of 4 of the bone metastasis genes among the SCPs used in this study (boxed and labeled by SCP, with the color of the label corresponding to tissue tropism). GAPDH, loading control.

metastasize to the lung. Shortly after tail vein injection, all detectable cells became trapped in the lung (Figure 6A). Within the first few days, there was a substantial attenuation of this signal. In SCP6 and SCP26, this attenuation continued over the ensuing weeks, suggesting that as in the bone, these SCPs were unable to efficiently survive and grow in the lung. The highly bone-metastatic populations SCP2 and SCP46 were also unable to grow in the lung but were able to survive over the course of several weeks, as shown by their persistent bioluminescence signal. In contrast, SCP3 and SCP28, and to a lesser extent SCP32 and SCP43, were able to grow in the lung. To confirm the presence of lung metastases, we per-

formed histological analysis. Immunohistochemistry with CD31, which is a marker for vascular endothelial cells, revealed multiple areas of perivascular tumor growth and growth within the capillary-rich lung parenchyma (Figure 6C).

It is hypothesized that growth at metastatic sites is enhanced by genes that confer productive tumor-stroma interaction. Thus, metastatic cells that grow well at one site may not grow well at another. Based on the metastatic tropisms of each SCP defined by BLI, we ranked SCPs according to their growth kinetics in either bone or lung. As shown in Figures 4B and 6B, SCPs with a higher rank order for bone were color-coded in red, and those with a higher rank order



To validate that metastasis-specific genes were among the 1,267 differentially expressed genes, we determined how many of the 102 unique genes from our previously described (12) and independently derived bone metastasis gene set (represented on the U133A GeneChip by 127 probe sets) were among the 1,267 genes. As seen in the Venn diagram in Figure 7C, 50 of the 127 bone metastasis genes were overlapping. This set of 50 included *IL11*, *CTGF*, and *CXCR4*, three genes that were determined to specifi-



cally cause bone metastasis (12). Accordingly, Figure 7D demonstrates that the expression of these genes strictly correlated with bone-specific growth (Figure 7D).

Segregation of primary tumors using a bone metastasis gene expression signature. The existence of a poor-prognosis gene expression signature from the bulk expression data of primary breast cancers suggests that the emergence of cells that express metastasis genes may occur early during tumorigenesis. Therefore, we wanted to determine whether the bone metastasis genes that we identified in our MDA-MB-231 model system in the mouse (12) could be detectable within primary breast carcinomas. To this end, we used the 50 bone metastasis genes expressed among the bone-metastatic SCPs. Hierarchical clustering of all 63 primary breast tumors in our cohort did not robustly distinguish those tumors that gave rise to bone metastasis from those that did not (data not shown). This suggests that either our bone metastasis signature carries little predictive value or our genes are expressed only by an undetectable subpopulation of tumor cells.

To help distinguish between these two possibilities, we restricted our analysis to those primary tumors that gave rise to distant metastasis (mainly to bone and/or to lung) (Figure 8). Under these conditions, the 50 bone metastasis genes could be used to divide the primary breast carcinoma groups into two major clusters with an overall reproducibility index (R index) of 0.90, which is indicative of the robustness of this cluster. The primary breast carcinomas that gave rise to bone metastasis were predominantly associated with the second cluster. In contrast, those samples that produced lung metastasis were mainly grouped together by the first cluster. The 50 bone metastasis genes were also clustered together into six groups based on similarity in their expression pattern. Gene cluster 2 represented genes that were generally upregulated in the primary tumors that developed bone metastasis. Genes in this cluster included *CTGF* and *IL11*, in addition to other genes that are upregulated in the bone-metastatic SCPs, including *NAP1L3*, *DUSP1*, *ADAMTS1*, and *SOCS2* (Supplemental Table 1). Some genes that are upregulated in the bone-metastatic SCPs, such as *MMP1*, are not selectively upregulated in the breast carcinoma primary tumors that develop bone metastasis; for example, *MMP1* is also involved in lung metastasis (our unpublished observations). The failure of other genes to display concordant expression patterns in the SCPs and the breast primary tumors may be because they are not biologically relevant and/or because of unknown peculiarities of the clinical data set or the MDA-MB-231 model system. Nonetheless, these data suggest that the development of distant sites of metastasis in breast cancer patients is related to differences in the gene expression pattern that is discernible by our bone metastasis gene expression signature.

Discussion

In this study, we have demonstrated that SCPs from a metastatic parental breast cancer population carry a poor-prognosis signature. This signature varied little from SCP to SCP; however, the metastatic activity of different SCPs varied significantly. With the sensitivity afforded by noninvasive BLI coupled with fluorescence microscopy, we were able to fully characterize the metastatic activities of individual SCPs by evaluating tissue tropism and growth kinetics. We determined that some SCPs were capable of efficient metastasis to bone, others metastasized better to lung, and a minority were also able to colonize and grow within the adrenal gland and/or other sites. This activity resembles the typical distribution of breast cancer metastases observed in patients. Some

SCPs exhibit multiple tropisms, while others, in contrast, are only weakly metastatic and/or give rise to dormant lesions. The presence of cells with different metastatic properties from the same pleural effusion-derived cell line may reflect an accumulation of circulating tumor cells from multiple metastatic sites within the pleural fluid of the patient from which the cells were derived.

Although we cannot rule out the possibility that minor variations in the poor-prognosis signature may contribute to these differences in metastatic phenotypes, hierarchical clustering based on the poor-prognosis genes does not clearly segregate the SCPs into different groups that correlate with particular aspects of metastatic activity such as colonization and growth within specific organs. This suggests that the genes that make up the poor-prognosis signature do not control these more specific metastatic properties. In contrast, hierarchical clustering based on the entire gene expression data set does segregate the SCPs into different groups with different organ tropisms. The poor-prognosis signature was defined in a way that does not take into account particular characteristics of metastasis such as tissue tropism and growth kinetics. In a recent report comparing human primary breast tumors to distant metastases from various organs, the primary tumor showed extensive genetic similarity to the distant metastasis from the same patient, and a "supervised" method was unable to generate a classifier to distinguish primary tumors from metastases (15). These results are in line with the concepts of a poor-prognosis signature; however, because the metastasis samples were from various organs, the presence of site-specific metastasis genes could not be determined. Thus, the poor-prognosis signature may be composed of gene expression events acquired early during primary tumor development that function to endow tumor cells with baseline metastatic properties or that mark a particular cell phenotype that is liable to express metastatic functions. Indeed, MBA-MD-231 cells are derived from the pleural effusion of a patient with widespread metastatic disease, and all of the individual clones from this population that we analyzed show at least some level of metastatic activity.

Based on the identification of metastasis genes associated with osteolytic bone metastasis, our previous study proposed that in addition to the poor-prognosis signature, metastatic cells need to acquire a genetic "tool box," or a set of genes that confer the functions necessary for efficient tissue-specific growth. The genes that make up this "tool box" would be regarded as metastasis-specific genes that are acquired through mutation or epigenetic changes. However, the classification of genes into this category would require a level of specificity such as tissue tropism. Our current study provides support for this requisite, as the expression of these genes strictly correlated with efficient bone metastasis and not with other recognizable aspects of metastatic activity. In addition, multidimensional scaling of genes that are differentially expressed across SCPs defines groups that correlate with primary tissue tropism, and our bone metastasis gene set overlaps with these differentially expressed genes. We expect that within these differentially expressed genes, a lung metastasis gene set will also exist (our unpublished observations). Thus, SCPs with different genetic profiles can exhibit marked differences in their ability to colonize and to grow exponentially in various metastatic sites. These results support the idea of the importance of productive tumor-stroma interactions that foster metastatic growth, consistent with Paget's "seed and soil" hypothesis (16), or interactions such as those between tumor and vasculature that result in differential tissue arrest.



Some of the SCPs that we analyzed demonstrated the ability to grow effectively at more than one metastatic site. For example, SCP28 grew well in both the bone and the lung, SCP3 was metastatic to both lung and adrenal, and SCP2 exhibited both bone and adrenal tropism. In contrast, SCP46 was metastatic only to the bone. The multi-tropic properties of metastatic cells raise the possibility that metastatic cells from one site may spawn metastasis to another site. Because there are limited clinical situations in which single metastasis or oligometastasis is effectively treated by surgical excision, knowledge of whether metastatic cells are single or multi-tropic may be of important clinical relevance.

A metastatic cell must complete a series of sequential steps in order to successfully colonize and grow at a distant site. Our data suggests that the expression of a poor-prognosis signature can mark only a baseline ability to accomplish some of these steps. The signature may comprise genes related to the early oncogenic changes that drive primary tumor formation, but is absent in genes that dictate organ-specific metastatic activity. These additional metastasis genes provide the capability to become fully metastatic and confer properties such as organotropism. It is unclear whether these metastasis genes are acquired during the growth of the primary tumor or during colonization at a distant site (17). Indeed, our hierarchical clustering of a mixed cohort of primary breast tumors with a bone metastasis gene expression signature (12) did not allow robust classification of those tumors that gave rise to bone metastasis versus those that did not. Nonetheless, this signature was able to distinguish between primary breast carcinomas that preferentially metastasized to bone from those that preferentially metastasized elsewhere. These results suggest that the development of distant sites of metastasis in breast cancer patients is related to differences in primary tumor gene expression pattern that are discernible by our bone metastasis gene expression signature. A further enrichment of the list of bone metastasis genes may allow in the future accurate prediction of the bone metastasis tropism of breast cancer primary tumors.

Methods

Cell culture and retroviral gene transfer. MDA-MB-231 cells were obtained from ATCC and were cultured in Dulbecco's modified Eagle's, high glucose supplemented with 10% FBS. SCPs were derived from MDA-MB-231 cells as described previously (12). The construction and retroviral gene transfer of the triple-modality reporter gene TGL has been described previously (14). In brief, 20 µg of the TGL reporter plasmid SFG-^{NES}TGL was transfected into the GPG29 packaging cell line with Lipofectamine 2000 (Invitrogen). Virus-containing supernatants were harvested between 72 and 96 hours, were filtered with a 0.45-µm syringe filter, and were used to infect MDA-MB-231 SCPs for 12–24 hours in the presence of 8 µg/ml of polybrene (Sigma-Aldrich). At 72 hours after infection, successful gene transfer was confirmed by visualization of GFP by fluorescence microscopy. These cells were enriched by fluorescence-activated cell sorting (FACS-Vantage; Becton Dickinson). Luciferase activity was confirmed in vitro by seeding of 1×10^5 cells into a 24-well plate followed by the addition of 0.03 mg of D-Luciferin (Xenogen). Luciferase activity was measured with the IVIS Imaging System (Xenogen).

Mouse xenografting. For intracardiac injections, subconfluent cells were harvested, washed in PBS, and resuspended at a concentration of 1×10^6 cells/ml. BALB/c nude mice (NCI) were anesthetized by intraperitoneal injection of ketamine (100 mg/kg) and xylazine (10 mg/kg) and were placed in the supine position. With a 26-gauge needle, 1×10^5 cells were injected into the left ventricle via the third intercostal space after visualization of arterial blood flow into the syringe. For tail vein injections, unanesthetized mice

were warmed with a heat lamp to allow for venous dilation. Mice were then placed into a plastic retraining apparatus, and 2×10^5 cells were injected via the lateral tail vein. Successful injections were confirmed by immediate BLI. All animal studies were performed in accordance with an IACUC-approved protocol at the Memorial Sloan-Kettering Cancer Center.

BLI and analysis. Anesthetized mice were injected retro-orbitally with 75 mg/kg of D-Luciferin (Xenogen) in PBS. Bioluminescence images were acquired with the IVIS Imaging System (Xenogen) at 2–5 minutes after injection. Acquisition times at the beginning of the time course started at 60 seconds and were reduced in accordance with signal strength to avoid saturation. Analysis was performed using LivingImage software (Xenogen) by measurement of photon flux (measured in photons/s/cm²/steradian) with a region of interest (ROI) drawn around the bioluminescence signal to be measured. For bone metastasis, an ROI was drawn around the major bioluminescence signal from the hindlimb, forelimb, or pelvis/sacrum. For lung metastasis, an ROI was used that encompassed the thorax of the mouse. For determination of the "fold increase" above background, average background measurements were obtained using the same ROI on a corresponding region from control mice. Data were divided by the average background measurement and were normalized to the signal obtained immediately after xenografting (day 0).

Histology. For whole-mount analysis, sacrificed mice were frozen in liquid nitrogen and were stored at -80°C. Prior to frozen sectioning, tissue was embedded in M1 embedding media (Shandon). Sections 20 µm in thickness were mounted on slides and were fixed with 100% methanol for 30 seconds. GFP was visualized in these mounted sections using a fluorescence microscope. H&E staining was then performed on serial sections of interest. For immunohistochemistry for CD31, lungs were fixed in 4% paraformaldehyde overnight and were incubated in 30% sucrose for an additional 12–24 hours prior to cryosectioning. CD31 staining was performed with the Discovery AutoStainer (Ventana Medical Systems) and anti-CD31 (sc-1506; Santa Cruz Biotechnology) at a concentration of 1 µg/ml.

DNA microarray analysis. Methods for RNA extraction, labeling, and hybridization for DNA microarray analysis of the cell lines have been described previously (12). For the primary breast tumor data, tissues from primary breast cancers were obtained from therapeutic procedures performed as part of routine clinical management. Samples were "snap-frozen" in liquid nitrogen and were stored at -80°C. Each sample was examined histologically with H&E-stained cryostat sections. Regions were manually dissected from the frozen block to provide a consistent tumor cell content of more than 70% in tissues used for analysis. All studies were conducted under protocols approved by the Memorial Sloan-Kettering Cancer Center Institutional Review Board. RNA was extracted from frozen tissues by homogenization in TRIzol reagent (GIBCO-BRL; Invitrogen Corp.) and was evaluated for integrity. Complementary DNA was synthesized from total RNA using a T7 promoter-tagged dT primer. RNA target was synthesized by in vitro transcription and was labeled with biotinylated nucleotides (Enzo Biochem). Labeled target was assessed by hybridization to Test3 arrays (Affymetrix).

All gene expression analysis was carried out using the Affymetrix U133A chip. Analysis of the poor-prognosis signature was performed using GeneSpring 6.1 (Silicon Genetics) with a list of genes from the 70 genes comprising the poor-prognosis signature that are present on the U133A chip. For multidimensional scaling and hierarchical clustering, Affymetrix data were imported into BRBArray Tools 3.1 (developed by Richard Simon and Amy Peng Lam; <http://linus.nci.nih.gov/BRB-ArrayTools.html>). Hierarchical clustering was performed using either Euclidean distance or Pearson correlation. Cluster reproducibility was reported as an R index (18). To obtain a list of genes that are broadly differentially expressed among the SCPs, we applied a filter to the 22,238 genes; this filter eliminated genes in



which expression levels differed by at least either 1.5-fold or twofold from the mean expression level in less than half of the data sets. An additional filter was applied to eliminate genes with an absent detection call in all of the datasets. The final filtered list comprised 1,267 genes (1.5-fold filter) or 286 genes (twofold filter). This list was used in both multidimensional scaling and hierarchical clustering. Other filtering criteria were also tested and gave comparable results.

CXCR4 staining for flow cytometry. Subconfluent cells were trypsinized and were washed twice in cold PBS. Phycoerythrin-conjugated anti-human CXCR4 (BD Pharmingen) or control IgG was incubated in FACS buffer (0.1% sodium azide and 1% bovine serum albumin in PBS) for 1 hour at 4°C. Cells were subsequently washed twice in PBS and, finally, were resuspended in FACS buffer. Cells were analyzed by flow cytometry using a BD FACSCalibur unit, and subsequent data analysis was done using FlowJo software.

Acknowledgments

We are indebted to Juri Gelovani for invaluable discussions; Weiping Shu for expert technical assistance; Agnes Viale and Julia Zhao of the Genomics Core Facility for microarray data analysis; and Katia Manova and the staff of the Molecular Cytol-

ogy Core Facility for assistance with immunohistochemistry. A.J. Minn is a recipient of the Leonard B. Holman Research Pathway fellowship, and Y. Kang is the recipient of a postdoctoral fellowship from the Irvington Institute for Immunological Research. G.P. Gupta is supported by the NIH Medical Scientist Training Program grant GM07739 and a fellowship from the Katherine Beineke Foundation. J. Massagué is an Investigator of the Howard Hughes Medical Institute. This research is also supported by NIH grant P01-CA94060 to J. Massagué and US Army Medical Research grant DAMD17-02-0484 to W.L. Gerald.

Received for publication June 2, 2004, and accepted in revised form November 2, 2004.

Address correspondence to: Joan Massagué, Cell Biology Program, Box 116, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. Phone: (212) 639-8975 Fax: (212) 717-3298; E-mail: j-massague@ski.mskcc.org.

Dilip D. Giri's present address is: Department of Pathology and Laboratory Medicine, Brown University, Providence, Rhode Island, USA.

1. Fidler, I.J. 2003. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat. Rev. Cancer*. **3**:453-458.
2. Chambers, A.F., Groom, A.C., and MacDonald, I.C. 2002. Dissemination and growth of cancer cells in metastatic sites. *Nat. Rev. Cancer*. **2**:563-572.
3. Tarin, D., et al. 1984. Clinicopathological observations on metastasis in man studied in patients treated with peritoneovenous shunts. *Br. Med. J. (Clin. Res. Ed.)*. **288**:749-751.
4. Tarin, D., Vass, A.C., Kettlewell, M.G., and Price, J.E. 1984. Absence of metastatic sequelae during long-term treatment of malignant ascites by peritoneovenous shunting. A clinico-pathological report. *Invasion Metastasis*. **4**:1-12.
5. Fidler, I.J. 1970. Metastasis: quantitative analysis of distribution and fate of tumor emboli labeled with 125 I-5-iodo-2'-deoxyuridine. *J. Natl. Cancer Inst.* **45**:773-782.
6. Cameron, M.D., et al. 2000. Temporal progression of metastasis in lung: cell survival, dormancy, and location dependence of metastatic inefficiency. *Cancer Res.* **60**:2541-2546.
7. van 't Veer, L.J., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. **415**:530-536.
8. van de Vijver, M.J., et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**:1999-2009.
9. Ramaswamy, S., Ross, K.N., Lander, E.S., and Golub, T.R. 2003. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**:49-54.
10. Bernards, R., and Weinberg, R.A. 2002. A progression puzzle. *Nature*. **418**:823.
11. Fidler, I.J., and Kripke, M.L. 2003. Genomic analysis of primary tumors does not address the prevalence of metastatic cells in the population [correspondence]. *Nat. Genet.* **34**:23; author reply, 25.
12. Kang, Y., et al. 2003. A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell*. **3**:537-549.
13. Cailleau, R., Young, R., Olive, M., and Reeves, W.J., Jr. 1974. Breast tumor cell lines from pleural effusions. *J. Natl. Cancer Inst.* **53**:661-674.
14. Ponomarev, V., et al. 2004. A novel triple-modality reporter gene for whole-body fluorescent, bioluminescent, and nuclear noninvasive imaging. *Eur. J. Nucl. Med. Mol. Imaging*. **31**:740-751.
15. Weigelt, B., et al. 2003. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc. Natl. Acad. Sci. U. S. A.* **100**:15901-15905.
16. Paget, S. 1989. The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev.* **8**:98-101.
17. Schmidt-Kittler, O., et al. 2003. From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **100**:7737-7742.
18. McShane, L.M., et al. 2002. Methods of assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*. **18**:1462-1469.

**Breast Cancer Bone Metastasis
Mediated by the Smad Tumor Suppressor Pathway**

**Yibin Kang^{1,5}, Wei He¹, Shaun Tulley¹, Gaorav P. Gupta¹,
Inna Serganova³, Chang-Rung Chen^{1,6}, Katia Manova-Todorova²,
Ronald Blasberg³, William L. Gerald⁴ and Joan Massagué¹**

¹ Cancer Biology and Genetics Program and Howard Hughes Medical Institute,

² Molecular Cytology Laboratory, and Departments of ³ Neurology and ⁴ Pathology

Memorial Sloan-Kettering Cancer Center, New York

⁵ Y.K current address: Department of Molecular Biology, Princeton University, Princeton, NJ 08544

⁶ C.R. current address: ArQule Biomedical Institute, 333 Providence Highway, Norwood, MA 02062

Correspondence:

Joan Massagué

Cancer Biology and Genetics Program, Box 116

Memorial Sloan-Kettering Cancer Center

1275 York Avenue, New York, NY10021, USA

Phone: 212-639-8975 Fax: 212-717-3298

j-massague@ski.mskcc.org

When breast cancer spreads to the bone, an osteolytic vicious cycle may arise whereby tumor cells instigate local osteoclasts to mobilize bone-derived TGF β that further activates the tumor¹. TGF β can signal by means of Smad transcription factors², which are quintessential tumor suppressors that inhibit cell proliferation^{3,4}, and by means of Smad-independent mechanisms which are implicated in tumor progression^{5,6}. Although Smad mutations disable this tumor suppressive pathway in certain cancers, breast cancer cells frequently evade the cytostatic action of TGF β while retaining Smad function^{3,4}. Here we show that breast cancer cells can use the Smad pathway to promote bone metastasis. Functional imaging and immunohistochemical analysis reveal the presence of active Smad signaling in mouse and human bone metastatic lesions. Smad signaling is shown to be essential for the induction of the bone metastasis gene *interleukin-11 (IL11)*, and to significantly contribute to the formation of osteolytic bone metastases. AP1 is a key participant in Smad-dependent transcriptional activation of *IL11* and its overexpression in bone metastatic cells. Our findings provide direct functional evidence for a switch of the Smad pathway, from tumor-suppressor to pro-metastatic, in the development of breast cancer bone metastasis.

TGF β plays a crucial role as a growth-inhibitory cytokine in many tissues^{3,4}. The cytostatic effect of TGF β is mediated by a serine/threonine kinase receptor complex that phosphorylates Smad2 and Smad3, which then translocate into the nucleus and bind Smad4 to generate transcriptional regulatory complexes². *SMAD4* (also known as *Deleted in Pancreatic Carcinoma locus 4, DPC4*) and, to a lesser extent, *SMAD2* suffer mutational inactivation in a proportion of pancreatic cancers and colon cancers^{3,4}. However, tumor cells that evade this anti-proliferative control by other mechanisms may display an altered sensitivity to TGF β and undergo tumorigenic progression in response to this cytokine^{3,4}. Patients whose pancreatic or colon tumors express TGF β receptors fare less well than those with low or absent TGF β receptor expression in the tumor⁷. In mouse models of breast cancer, TGF β signaling promotes lung^{8,9} and bone metastasis¹⁰. Although the tumorigenic actions of TGF β have been ascribed to Smad-independent

mechanisms⁶, we investigated whether the Smad pathway mediates bone metastasis in breast cancer.

Receptor-mediated phosphorylation of Smad2 at the C-terminus and accumulation of phospho-Smad2 in the nucleus are typical indicators of TGF β stimulation². To determine whether this pathway is active in bone metastasis, metastatic tissues from breast cancer patients were subjected to immunohistochemistry with anti-phosphopeptide antibodies against receptor-phosphorylated Smad2. Bone metastasis tissues from 16 breast cancer patients were obtained from therapeutic procedures performed as part of routine clinical management of these patients at our institution. Twelve of these samples showed prominent anti-phospho-Smad2 staining (Figure 1a), and this staining was concentrated in the nucleus (Figure 1b-e). Nuclear phospho-Smad2 staining was present both in the tumor cells and cells of the surrounding stroma (e.g. Figure 1b), suggesting that the entire field was under TGF β stimulation in these lesions. The other four metastasis samples analyzed showed little or no staining. Thus, a majority of breast cancer bone metastases exhibited evidence of Smad pathway activation.

Prompted by these results, we sought evidence for Smad-dependent transcriptional activity in bone metastasis by functional imaging in a mouse xenograft model. This model is based on the MDA-MB-231 cell line, which was derived from the pleural effusions of a breast cancer patient with metastatic disease¹¹. From parental MDA-MB-231 cells we isolated various sub-lines with distinct organ-specific metastatic behavior^{12,13}. The sub-line SCP2 is highly metastatic to bone via arterial circulation whereas sub-line SCP3 is highly metastatic to the adrenal glands. A retroviral reporter vector Cis-TGF β 1-Smads-HSV1-tk/GFP was created in which a fusion protein containing HSV1 thymidine kinase (HSV1-tk) and green fluorescent protein (GFP) was placed under the transcriptional control of a TGF β -responsive promoter element (Figure 2a). We chose the TGF β responsive element (T β RE) from the mouse germline *Ig α* promoter^{14,15}. This T β RE is recognized by Smad2/3-Smad4 in complex with RUNX family members and responds to TGF β in many different cell lines^{14,15}. RUNX activity in breast cancer cells is implicated in osteolytic bone metastasis¹⁶. Cis-TGF β 1-Smads-HSV1-tk/GFP was transduced into SCP2 and SCP3 cells together with a second retroviral vector SFG-

tdRFP-cmvFLuc expressing red fluorescent protein (tdRFP) ¹⁷ and firefly luciferase (Fluc) under constitutive promoters (Figure 2a). The RFP-positive cells expressed green fluorescence in response to TGF β , demonstrating responsiveness of the HSV1-tk/GFP construct (Figure 2b, c). When inoculated into the arterial circulation of immunodeficient mice, SCP2 cells formed aggressive bone metastases, as visualized by luciferase bioluminescence imaging (Figure 2d). These lesions also expressed TK activity, as determined by micro-positron emission tomography (micro-PET) (Figure 2d). SCP3 cells formed small bone metastases and very large adrenal metastases (Figure 2e top) ¹². Interestingly, while the small bone metastases formed by SCP3 expressed TK activity in the live animals, the large adrenal metastases formed by the same cells did not (Figure 2e top). The location of these lesions was verified by ex vivo bioluminescence of the affected organs after necropsy (Figure 2e, bottom). These results suggest that breast cancer cells undergo Smad-dependent transcriptional activation in the bone microenvironment.

We recently identified a set of genes that mediate osteolytic bone metastasis by MDA-MB-231 cells ¹². Among these genes, *IL11* was of interest because of its role as an enhancer of osteoclast differentiation ¹⁸ and as a mediator of osteolysis in breast cancer bone metastasis ^{19,20}. Enforced expression of *IL11* in MDA-MB-231 cells increases their bone metastatic activity ¹². Intriguingly, *IL11* is a TGF β inducible gene ^{12,21}, providing a mechanism for the pro-metastatic activity of TGF β in breast cancer. MDA-MB-231 cells are defective in TGF β cytostatic gene responses, including repression of *c-myc* and *Id* genes ²², but retain many responses that are common among normal epithelial cells ²³, including *IL11* induction (Figure 3a and Supplemental Table 1). A comparison of the basal expression of TGF β responsive genes in various MDA-MB-231 derivatives revealed a sharp (>9-fold) and selective increase in the basal expression of *IL11* in highly bone-metastatic sub-lines compared to the poorly metastatic sub-lines, and compared also to all the other TGF β responsive genes (Supplementary Table 2; summarized in Figure 3a). A smaller increase was observed in the basal expression of *CTGF*, which is another TGF β responsive gene implicated in bone metastasis ¹² (Supplementary Table 2). Thus the bone metastatic cells overexpressed certain TGF β

responsive genes that, in the context of the bone marrow microenvironment, stimulate osteolytic metastasis.

Several results suggested that *IL11* is an immediate TGF β target gene. *IL11* induction by TGF β is rapid, peaking at 2h and gradually declining thereafter (Figure 3b), and the protein synthesis inhibitor cycloheximide does not block this response (data not shown). TGF β stimulation induces the binding of Smad2/3 and Smad4 to the *IL11* promoter in chromatin immunoprecipitation experiments ¹². To determine whether the Smad pathway is required for *IL11* induction and bone metastasis, we analyzed MDA-MB-231 single cell progeny (SCP) sub-lines that were depleted of Smad4 by means of RNAi. Compared to parental cells or in vivo selected bone-metastatic populations, which are heterogeneous, SCPs are derived from single cells and, therefore, are more homogenous in genetic makeup ^{12,24}. Three bone metastatic sub-lines, SCP2, SCP25 and SCP28, were engineered to stably express the short-hairpin RNA (shRNA) probes Smad4-shRNA1 or Smad4-shRNA2, which target different regions of the *Smad4* mRNA. Expression of Smad4-shRNA1 reduced Smad4 protein levels by 70-90% in all three SCPs whereas Smad4-shRNA2 almost completely eliminated Smad4 production (Figure 3c). As a control, we engineered a Smad4 vector (pBabe-hygro-Flag-Smad4M) containing two silent mutations in the sequence targeted by Smad4-shRNA1 and an N-terminal flag epitope distinguishing the exogenous product from endogenous Smad4. Transduction of this retrovirus ensured expression of Smad4 in cells containing Smad4-shRNA1 (Figure 3c). As determined by Northern blot analysis, the *IL11* response to TGF β was very weak in cells expressing Smad4-shRNA1 and undetectable in cells expressing Smad4-shRNA2 (Figure 3d for SCP25; data not shown for SCP2 and SCP28). Expression of Smad4M restored the TGF β response in Smad4-shRNA1 expressing cells. A similar response pattern was observed at the level of IL11 protein secretion, as determined by ELISA (Figure 3e; and data not shown for SCP2 and SCP28). Thus, Smad4 is essential for TGF β activation of *IL11* expression.

To further investigate the role of Smad factors in the *IL11* response to TGF β , we focused on a 100bp region immediately upstream of the TATA box in the *IL11* promoter. This region mediates the TGF β response of the *IL11* promoter in human epithelial and carcinoma cells ^{21,25}. A reporter construct under the control of the minimal *IL11* promoter

[pIL11-(100)-Luc]²¹ was unresponsive to TGF β in the Smad4-deficient breast cancer cell line MDA-MB-468²⁶ (Figure 4a). Expression of exogenous Smad4 enabled TGF β induction of this promoter, and this effect was further enhanced by co-transfection of Smad2 or Smad3 (Figure 4a), arguing that Smads mediate transcriptional activation from this promoter region.

This 100bp region includes two AP1 binding sites, which are critical for *IL11* transcription^{21,25}, and an adjacent GC-rich (92% GC) sequence with two putative SP1 sites (Figure 4b). No canonical Smad binding element (AGAC sequence) is present in this region. However, Smads can bind to GC-rich sequences in certain promoters². Deletion analysis of the *IL11* promoter region by means of a reporter construct indicated that the response to TGF β minimally requires the 5' AP1 site and an adjacent GC-rich sequence (Figure 4b). In electrophoretic mobility shift assays, recombinant Smad4 bound to the wild type minimal *IL11* promoter probe, resulting in the formation of a complex that could be shifted by addition of anti-Smad4 monoclonal antibody (Figure 4c). Mutation or deletion of the AP1 sites decreased but did not abolish Smad4 binding to the probe, whereas the AP1 sites alone did not bind Smad4 (Figure 4c). The binding of endogenous Smad and AP1 factors to this region was assessed by means of oligonucleotide precipitation assays. MDA-MB-231 cells were incubated with or without TGF β for 2h, lysed, and precipitated with biotinylated double-stranded DNA probes. Immunoblotting of DNA-bound factors demonstrated TGF β -dependent binding of endogenous Smad3 and Smad4 to the wild type *IL11* minimal promoter region, and TGF β -independent binding of the endogenous AP1 component JunB to this region (Figure 4d). Deletion or mutation of the AP1 sites eliminated binding of JunB and weakened Smad binding.

Consistent with a role of AP1 in the *IL11* response to TGF β in the breast cancer cells, the AP1 activator 12-O-tetradecanoylphorbol-13-acetate (TPA)²⁷ increased the basal level of *IL11* expression as well as the level upon TGF β stimulation, whereas the AP1 inhibitor curcumin²⁷ abolished the activation of *IL11* by TGF β (Figure 4e). As determined using an AP1 reporter construct (4xAP1-*luciferase*), the basal level of AP1 activity was significantly higher in the highly metastatic sub-lines SCP2, SCP25, SCP28 and 1833 than in poorly metastatic sub-lines SCP4 and SCP6 or parental MDA-MB-231

cells (Figure 4f). The level of AP1 activity in these cell populations was closely correlated with the basal level of *IL11* expression (Figure 4f; refer to Supplementary Table 2). No change in 4xAP1 luciferase activity was observed after 4 h of TGF β treatment (data not shown). Collectively, these results suggest that TGF β -activated Smad proteins bind to the GC-rich region in the proximal *IL11* promoter. This binding is strengthened by the presence of a proximal AP1 site, and transcriptional activation results from a cooperation between Smad3 and AP1. These observations also indicate a role of AP1 in the hyperactivity of *IL11* in bone metastatic MDA-MB-231 cells.

Having shown that the TGF β response of a bone metastasis gene in these cells required Smad function, we tested the contribution of Smad signaling to the metastatic process itself. Wild-type, Smad4-knockdown, and Flag-Smad4M versions of the various SCPs were infected with a retroviral vector expressing HVS1-tk/GFP/luciferase triple fusion protein²⁸. The cells were inoculated into the left cardiac ventricle of immunodeficient mice to allow the formation of bone metastasis. As determined by bioluminescence imaging of luciferase activity, the inoculated cells became immediately distributed throughout the entire animal followed by extensive clearing within one week (Figure 5a). Accumulation of luciferase signal was clear 14 days after injection and became more intense over the following weeks. To quantify the rate of metastatic growth in bone, a region of interest (ROI) was drawn around the bone metastases signals near the joint of the affected hind limbs, and the normalized photon counts of each metastasis was plotted (Figure 5b). A linear correlation between the intensity of the bioluminescence and tumor burden is obtained using this method²⁹. Suppression of Smad4 activity by two different shRNA constructs caused a significant reduction in the growth rate of bone metastatic lesions (Figure 5a, b). Restoration of Smad4 function by the shRNA-insensitive Smad4M construct restored the wild-type rate of metastatic growth (Figure 5a, b). These results were consistently observed in all three SCPs tested (Figure 3c, and data not shown for SCP2 and SCP28). Formation of overt osteolytic bone metastases was monitored by weekly full-body x-ray imaging of the mice. Smad4 depletion consistently reduced the rate of bone metastasis formation in all three MDA-MB-231 SCPs and in the in vivo-selected bone-metastatic population 1833¹² (Figure 5c). A significant level of metastatic activity still remained after Smad4

depletion, which is consistent with the TGF β -independent involvement of several genes (*MMP1*, *CXCR4*, *Osteopontin* and others) in these lesions ¹². Smad4-knockdown did not decrease the growth rate of the SCPs or 1833 cells in culture (data not shown) or their ability to form subcutaneous tumors in mice (Figure 5d), arguing that the Smad4-dependent growth of these tumors is specifically stimulated by the bone microenvironment.

In sum, our results show that the Smad tumor suppressor pathway may become pro-metastatic in breast cancer. The intrinsic genomic instability of tumor cell populations allows for the selection of functions that favor growth in a given environment. Thus, a bone metastatic lesion will harbor functions that the bone environment selects for. We speculate that pro-metastatic Smad-mediated gene responses can emerge once this pathway becomes uncoupled from tumor-suppressor effects. If at that point a Smad pathway can provide metastatic functions to cancer cells, it likely will be selected as a pro-metastatic force. Smad-responsive genes like *IL11* and others can provide an advantage to cancer cells in a TGF β -rich bone microenvironment. Therefore, an increase in the basal expression of these genes coupled with their further induction by bone-derived TGF β would favor tumor growth in the bone. Our results are fully consistent with this possibility. By implicating the Smad pathway in the osteolytic vicious cycle of breast cancer metastasis ¹, our results additionally call attention to the possibility of therapeutically targeting this pathway ^{6,30} in TGF β -rich metastatic sites.

Experimental procedures

Tumor sample analysis.

Formalin-fixed paraffin embedded (FFPE) bone metastasis tissues were obtained from therapeutic procedures performed as part of routine clinical management of breast cancer patients at our institution. Hematoxylin and eosin stained sections were examined for regions that contained both tumor cells and stroma, which were further analyzed for phosphorylated Smad2 on serial sections. All studies were conducted under MSKCC Institutional Review Board approved protocols.

Immunohistochemistry

Immunohistochemical analysis was performed with a Discovery XT System (Ventana Medical Systems) using tissue sections blocked for 30 minutes in 10% normal goat serum (Vector Laboratories; catalog# S-1000) and 2% BSA. Incubation with anti-phospho-Smad2 (Ser465/467) primary antibody (Cell Signaling; catalog#3101; dilution 1:500) was carried out for 3 h at room temperature followed by a 1 h incubation with biotinylated anti-rabbit secondary antibody at 1:200 dilution (Vectastain ABC Kit Rabbit IgG catalog# PK-6101) and DAB detection kit (Ventana Medical Systems) according to the manufacturer instructions.

TGF β 1-Smads-HSV1-tk/GFP reporter system

Double-stranded complementary oligonucleotides, containing a sequence from the mouse germline Ig α promoter 5'-AATTCGGCCATGTGGTCAGACACACCTGTCTCCACCACAGCCAGACCACAGGCCAGACATGACGTGGAGGTT-3'³¹, were used to construct the TGF β 1-Smads-HSV1-tk/GFP reporter vector. After annealing of oligonucleotides, the resulting DNA fragment was cloned into the *EcoRI* and *XbaI* sites of the dxNFAT-tk/GFP-Neo vector³² in place of the NFAT enhancer element. Thus, the Herpes Simplex Virus 1 thymidine kinase-eGFP (HSV1-tk/GFP) fusion reporter gene was linked to the enhancer elements specific for Smad-AML transcriptional complexes. The resulting plasmid was transfected into the GPG29 packaging cell line with Lipofectamine2000 (Invitrogen, Carlsbad, CA). The retrovirus-containing medium was

collected for 4 consecutive days and stored at -80°C . The retrovirus was then used to transduce MDA-MB-231 cells and their sub-line SCP3^{12,24}. Selection of stable transfectants was accomplished by adding 1 g/L of G418. Cells containing the TGF β 1-Smads-HSV1-tk/GFP reporter system were further transduced with a second retroviral vector SFG-tdRFP-cmvFLuc, in which tdRFP¹⁷ and firefly luciferase encoding cDNAs were placed under constitutive promoters. RFP positive cells were sorted by FACS.

The retrovirus vector encoding a TK-eGFP-Luciferase triple fusion proteins has been previously described²⁸.

Transcriptomic profiling and clustering analyses

Tissue collection, RNA sample collection and generation of biotinylated complementary RNA (cRNA) probe were carried out essentially as described in the standard Affymetrix (Santa Clara, CA) GeneChip protocol. Each sample was hybridized with an Affymetrix Human Genome U133A microarray for 16h at 45°C . Absolute analysis of each chip and comparative analysis of TGF β treated samples with the untreated samples were carried out using the Affymetrix Microarray Suite 5.0 Software. Genes whose expression level was changed by more than two fold with $p < 0.001$ were scored as TGF β regulated genes. Dendrogram illustration of TGF β gene responses in MDA-MB-231 and MCF-10A cell lines were produced using GeneSpring (Silicon Genetics, CA) software.

Cell culture and retroviral transduction

Parental (ATCC) MDA-MB-231 cell line and its various sublines, as well as A549 cell line were maintained in DMEM medium supplemented with 10% fetal bovine serum (FBS), penicillin, streptomycin and fungizone. Phoenix cells, a helper cell line for retrovirus production, were maintained in DMEM medium supplemented with 10% fetal bovine serum (FBS), 1% glutamine and antibiotics.

Retroviruses expressing Smad4 shRNA, FLAG-Smad4M, or imaging proteins, were produced from amphotropic Phoenix packaging cell line. Phoenix cell transfections were performed using LipofectAMINE (Invitrogen), according to the manufacturer's instructions. Viruses were harvested 48h and 72h after transfection, filtered, and used to

infect MDA-MB-231 cell cultures in the presence of 5µg/ml of polybrene. Infected cells were selected by fluorescence activated cell sorting (FACS) for GFP positive cells, or by selection for puromycin or hygromycin resistance. To avoid clonal variations, we pooled at least 2000 individual transfectants for each stable cell line produced by transduction.

Plasmids

The minimal *IL11* promoter region containing the TATA box (-31 to +52) was cloned as a KpnI/BglI fragment into the corresponding sites in the pXP2-luc (ATCC) to create pIL11-TATA-Luc. Various *IL11* promoter regions (Figure 1C) immediately upstream of the TATA box were then inserted as BamHI/KpnI fragments upstream of the TATA box to generate a series of luciferase reporters controlled by different regions of the *IL11* promoter. Retroviral vectors that encode shRNAs against *hSmad4* transcript were generated by cloning suitable oligonucleotide sequences into the pSUPER-retro-puro vector³³. The coding strand of the Smad4 targeting shRNAs were GGATGAATATGTGCATGAC (Smad4-shRNA1) and GGTGTGCAGTTGGAATGTA (Smad4 shRNA2). A cDNA sequence encoding FLAG epitope-tagged *hSmad4* was cloned to the BamHI/Sall sites of pBabe-hygro³⁴ to generate pBabe-hygro-Flag-Smad4. Silent mutations were generated by site-directed mutagenesis in the coding sequence of Tyr162 (from TAT to TAC) and Val163 (from GTG to GTT) to create a shRNA-insensitive version of Smad4 expression plasmid pBabe-hygro-Flag-Smad4M.

Luciferase reporter assays

Luciferase reporter assays were performed as previously described²². 100 pM TGFβ1 (R&D Systems), 10 µg/ml of cycloheximide (Sigma), 100nM TPA (Sigma), 70µM Curcumin (Sigma) were used to treat cells in various assays. Northern blot analysis was carried out as previously described²².

ELISA analysis

The production and secretion of IL11 in various sublines of MDA-MB-231 were determined in 24h-conditioned media using commercially available IL11 (R&D Systems) ELISA kits according to the manufacturer's instructions.

Electrophoretic mobility shift assay

Purified full-length Smad4 protein was used in this experiment. Complementary oligonucleotides corresponding to the wild-type *IL11* promoter and its mutants were annealed and end labeled with $\gamma^{32}\text{P}$ -ATP. The sequences for the probes are: 5'-GGGTGAGTCAGGATGTGTCAGGCCGGCCCTCCCCTGCCGCCTGCCCCCGCCCCGCCCCGCCAGGCCCC-3' for W.T., 5'-GGGTTCTTCAGGATTGTTCAGGCCGGCCCCCTCCCCTGCCGCCTGCCCCCGCCCCGCCAGGCCCC-3' for mAP1; 5'-GGC CGGCCCTCCCCTGCCGCCTGCCCCCGCCCCGCCAGGCCCC-3' for GC; 5'-GGGTGAGTCAGGATGTGTCA-3' for AP1 and 5'-GTAAGCCCGGCCAGCCGACCGGGGC3' for β -actin.

The DNA-protein binding reactions were performed and analyzed on a 5% nondenaturing gel (Brunet et al.,1999). For supershift assessment, 1ul of mouse monoclonal antibody (BD transduction Laboratories; Catalog # 610843) against Smad4 was preincubated with full-length recombinant His-Smad4 for 5 minutes on ice. DNA-protein complexes were visualized by autoradiography.

DNA precipitation assay

DNA precipitation assays were carried out as described previously³⁵. The sequences of oligonucleotides used are as follows: 5'-GGGTGAGTCAGGATGTGTCAGGCCGGCCCTCCCCTGCCGCCTGCCCCCGCCCCGCCAGGCCCA-3' for WT, 5'-GGGACAATCCGGACAATCCGGCCGGCCCTCCCCTGCC GCCTGCCCCCGCCCCGCCAGGCCCA-3' for mAP1, 5'-GGCCGGCCCTCCCCTGCCGCCTGCCCCCGCCCCGCCAGGCCCA-3' for GC. 5'-GGGTGAGTCAGGATGTGTCAGGCCGGCCCTCCCCTGCCGCC-3' for AP1GC5', and 5'-GGGTGAGTCAGGATGTGTCATGCCCCCGCCCCGCCAGGCCCA-3' for AP1GC3'. Nucleotides that were

mutated in the corresponding mutant probes were highlighted with bolded, underlined letters. The sequence for TIE has been reported previously ³⁵.

Intracardiac injections

Cells were harvested from subconfluent cell culture plates, washed with PBS, and resuspended at 10^6 /ml concentrated in PBS. 0.1ml of the suspended cells were injected into the left cardiac ventricle of 4 week old, female BALB/c-nu/nu nude mice (NCI) using 26 gauge needles as previously described ¹⁰. Mice were anesthetized with ketamine (100mg/kg body weight) and xylazine (10mg/kg body weight) before injection. A successful injection was characterized by the pumping of arterial blood into the syringe and by immediate bioluminescence imaging.

Bioluminescence imaging and analysis

Anesthetized mice were retro-orbitally injected with 75 mg/kg of D-Luciferin (Xenogen) in PBS. Bioluminescence images were acquired using the IVIS Imaging System (Xenogen) at 2-5 minutes post-injection. Acquisition times at the beginning of the time course started at 60 seconds and were reduced in accord with signal strength to avoid saturation. Analysis was performed using LivingImage software (Xenogen) by measuring photon flux (measured in photons/sec/cm²/steradian) using a region of interest (ROI) drawn around the bioluminescence signal to be measured. Images were set at the indicated pseudo-color scale to show relative bioluminescent changes over time. Data were normalized to the signal obtained right after xenografting (day 0).

Micro-PET imaging

Micro-PET imaging was performed using ¹⁸F-2'-fluoro-2'-deoxy-1 β -D-arabionofuranosyl-5-ethyl-uracil ([¹⁸F]FEAU) as the HSV1-TK substrate, as previously described ³⁶. Two hours before whole body positron emission tomography (PET), the mice were administered [¹⁸F]FEAU (i.v. 100 μ Ci/animal). Imaging was performed on a microPET (Concorde Microsystems, Knoxville, TN) and images were acquired over 15 minutes under inhalation anesthesia (Isoflurane 2%).

Radiographic analysis of bone metastasis

Development of bone metastases was monitored by X-ray radiography. Mice were anesthetized, arranged in prone position on single-wrapped films (X-OMAT AR, Eastman Kodak, Rochester, NY), and exposed to an X-ray at 35kV for 15 seconds using a Faxitron instrument (Model MX-20; Faxitron Corp. Buffalo, IL, USA). Films were developed using a Konica SRX-101A processor and inspected for visible bone lesions.

Acknowledgements

We would like to thank Y-C Yang for reagents, A. Minn and L. Jayaraman for advice. We acknowledge the use of the Genomics Core Facility and the Flow Cytometry Core Facility at MSKCC. This research is supported by the W.M. Keck Foundation and NIH grant P01-CA94060 (JM), NIH grant P50-CA86438 (RB) and U.S. Army Medical Research grant DAMD17-02-0484 (WG). Y.K. was the recipient of a postdoctoral fellowship from the Irvington Institute for Immunological Research. G.P.G. was supported by the NIH Medical Scientist Training Program grant GM07739, a fellowship from the Katherine Beineke Foundation, and the Department of Defense Breast Cancer Research Program pre-doctoral traineeship award W81XWH-04-1-0334. J.M. is an Investigator of the Howard Hughes Medical Institute.

Figure Legends

Figure 1. Activated Smad pathway in breast cancer bone metastasis. **a**, Summary of phospho-Smad2 immunoreactivity in tumor cells and stromal cells in 16 samples of human breast cancer bone metastases; -, +, ++, +++ indicate none, weak, moderate and intense immunoreactivity respectively. **b-d**, Examples of intense immunohistochemical staining of receptor-phosphorylated Smad2 in breast cancer bone metastasis samples from different patients. The samples shown were chosen to illustrate the nuclear phospho-Smad2 staining in a metastatic island and the surrounding stroma (b), in a cluster of metastatic islands (c), or in a contiguous metastatic mass (d), as well as a cluster of islands stained using normal rabbit serum as a negative control.

Figure 2. Functional imaging of Smad signaling in breast cancer bone metastasis
a, Schematic representation of the retroviral vectors SFG-tdRFP-cmvFLuc, constitutively expressing tdRFP and firefly luciferase; and Cis-TGF β 1-Smads-HSV1-tk/GFP, expressing HSV-tk/GFP fusion protein in response to TGF β . **b** and **c**, SCP3 transduced with these two vectors were treated with TGF β or no additions for 24 h and analyzed by fluorescence microscopy (b) or two-color FACS (c). The constitutive tdRFP fluorescence is shown on the ordinate, and the HSV-TK/GFP fusion fluorescence, inducible by TGF β , is shown on the abscissa. **d** and **e**, *In vivo* bioluminescence and microPET imaging of metastases in mice. SCP2 (d) and SCP3 cells (e), bearing the SFG-tdRFP-cmvFLuc and Cis-TGF β 1-Smads-HSV1-tk/GFP vectors, were injected into the left cardiac ventricle and analyzed after 4 weeks (SCP2) or 18 weeks (SCP3). Bioluminescence imaging shows sites of metastases in the skull (in d, e) and adrenal gland (in e). [^{18}F]FEAU micro-PET images of tk/GFP reporter activation shows localization of radioactivity to the skull in both coronal and sagittal image planes. No visualization of the adrenal metastasis was seen on microPET imaging. Note non-specific accumulation of the tracer in the gastrointestinal tract and bladder attributable to clearance of the tracer. At necropsy, the head showing the skull and the adrenal metastasis plus kidney were removed and imaged ex vivo for photographic (-) and bioluminescence (+) imaging (e, lower panel).

Figure 3. Smad4-dependent transcriptional activation of *IL11* by TGF β . **a**, Basal expression levels of 50 TGF β -activated genes and 21 TGF β -repressed genes in MCF-10A and MDA-MB-231 cells were normalized to the same level. Responses of these genes to TGF β in each cell line were represented by different shades of red (degrees of activation) or blue (degrees of repression) in the dendrogram. The ratio of basal expression levels of these 71 genes in highly metastatic versus weakly metastatic MDA-MB-231 cells were represented by a bar graph in the right panel. **b**, Parental MDA-MB-231 cells were incubated with TGF β for the indicated times. Total RNA was subjected to Northern blot analysis using *IL11* and *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*) probes. **c**, Several single cell progenies (SCPs) derived from MDA-MB-231 were infected with retroviruses expressing Smad4-targeting shRNAs or shRNA-insensitive Flag-tagged Smad4. Protein expression was assessed by direct immunoblotting of total lysates using the indicated antibodies. **d**, SCP25 and its derivatives (refer to Figure 3c) were incubated in the absence or presence of TGF β for 2h. Total RNA was subjected to Northern blot analysis with indicated probes. **e**, SCP25 and its derivatives were treated with or without TGF β for 24 h. *IL11* production in the media was determined using an ELISA assay. Data are the average of triplicate determinations \pm S.D.

Figure 4. Role of AP1 and Smad in the basal activity and the TGF β response of the *IL11* promoter. **a**, Smad4-deficient MDA-MB-468 cells were transfected with 1 μ g of pIL11(-100)-Luc reporter plasmid ²¹, together with 0.5 μ g of the indicated Smad expression plasmids ³⁵, treated with or without TGF β , and analyzed for luciferase activity. Data are the average of triplicate determinations \pm S.D. **b**, Top: Nucleotide sequence of the minimal TGF β responsive region of the *IL11* promoter. Nucleotide sequence positions are indicated relative to the transcription start site. Two AP1 sites (red boxes) and a GC-rich sequence (green) containing two SP1 site (green boxes) are indicated. Bottom: A549 and MDA-MB-231 cells were transfected with the indicated *IL11* reporter constructs, treated with or without TGF β for 16-20 h prior to lysis, and analyzed for luciferase activity. The schematic representation of each promoter construct is shown on the left. Data are the average of triplicate determinations \pm S.D. **c**, γ ³²P-ATP end-labeled probes matching to the wild-type *IL11* proximal promoter region,

this region with mutant AP1 sites, or the indicated fragments of this region, were subjected to electrophoretic mobility shift analysis with recombinant full-length His-Smad4 protein. Antibody against Smad4 was added as indicated to create super-shifts. The β -actin promoter was used as a negative control. Schematic representations of the probes are shown at the top. **d**, MDA-MB-231 cells were incubated in the absence or presence of TGF β for 2h. Cell lysates were incubated with biotinylated oligonucleotides corresponding to the indicated *IL11* promoter probes. DNA-bound proteins were precipitated by streptavidin-agarose and detected by immunoblotting. A mutant *c-myc* TGF β response element (mTIE) was used as a negative control. **e**, A549 cells were incubated with 100nM TPA, 70 μ M curcumin or no additions for 30 minutes, and then with 100pM TGF β for the indicated period. Total RNA was subjected to Northern blot analysis with the indicated probes. **f**, Various MDA-MB-231 sublines were transfected with 1 μ g of 4xAP1-Luc reporter plasmid, and analyzed for luciferase activity 2d after transfection. Data are the average of triplicate determinations \pm S.D. The absolute values of *IL11* mRNA level as detected by Affymetrix U133A GeneChip were plotted in the same graph (yellow circles). The scales for the luciferase activity and for *IL11* GeneChip expression values were shown in the left and right sides of the graph, respectively.

Figure 5. Smad4 mediation of breast cancer bone metastasis. Wild-type and genetically modified SCP25 was labeled with the TGL reporter and 1×10^5 cells were injected into the left cardiac ventricle of five mice for each cell line. At the indicated days post-xenografting, bioluminescence images were acquired and quantified. **a**, Representative mice from each group are shown in the supine position. The intensity of the signal from days 24 and 36 are on equivalent scales, while day 0, 7 and day 14 are each on separate scales due to increasing signal strength and to avoid signal saturation. The normalized photon counts from the bone metastases in the hindlimbs were measured over the indicated time course and shown in **b**. **c**, Kaplan-Meier curves showing the incidence of bone metastasis by indicated wild-type and Smad4-knockdown MDA-MB-231 sub-lines. 10^5 tumor cells were inoculated into the left cardiac ventricle of nude mice. Metastasis was scored as the time to first appearance of a visible bone lesion by X-ray imaging of the whole mouse. The percent of animals in

each group that were free of detectable bone metastases is plotted. **d**, 10^6 tumor cells were injected subcutaneously into nude mice. Subcutaneous tumor growth was monitored and quantified by caliper measurements. No significant difference was found between wild-type and Smad4-knowdown cells.

Supplementary Table 1. Summary of TGF β target genes in three normal human epithelial cell lines and MDA-MB-231 breast cancer cells.

Shown are gene responses observed in at least two out of the three cell lines derived from normal tissues (HaCaT keratinocytes, MCF10A mammary epithelial cells and HPL1 lung epithelial cells) ²³, and the response of these genes in MDA-MB-231 cells. I: signal increased by TGF β by more than 2-fold; D: signal decreased by TGF β by more than 2-fold.

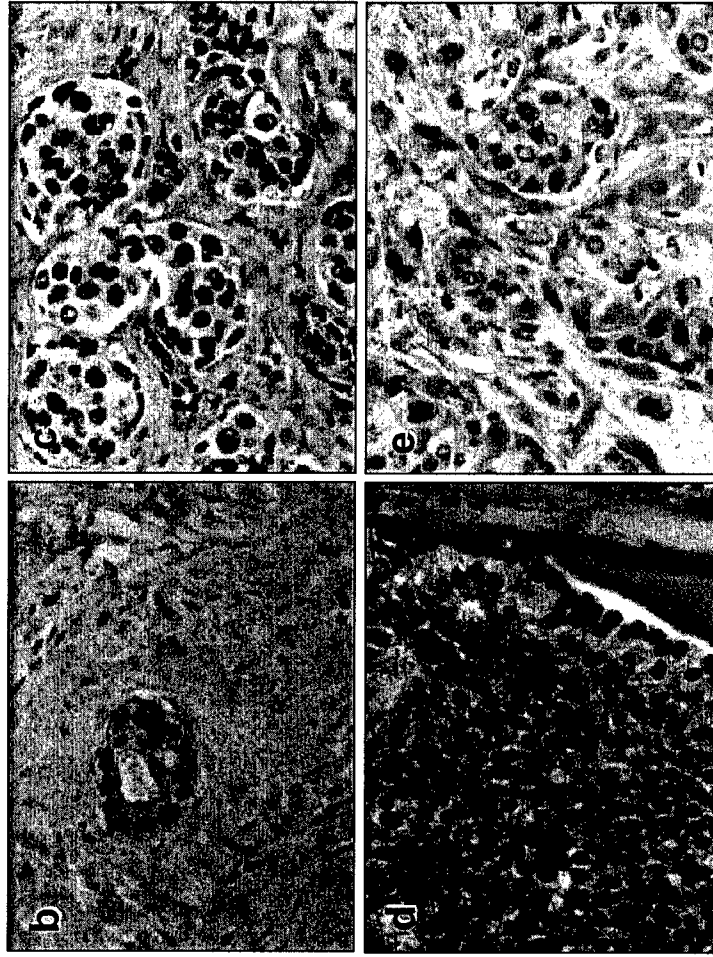
Supplementary Table 2. Basal level of TGF β epithelial cell target genes in various MDA-MB-231 sub-lines of different bone metastatic activity.

The list of genes examined corresponds to genes whose expression was increased or decreased in response to TGF β in at least two out of three cell lines (HaCaT keratinocytes, MCF10A mammary epithelial cells and HPL1 lung epithelial cells) derived from normal tissue (*3E TGF β response signature*) ²³. The two genes whose basal expression level was >3-fold higher in highly bone-metastatic cells compared to poorly metastatic cells are highlighted.

References

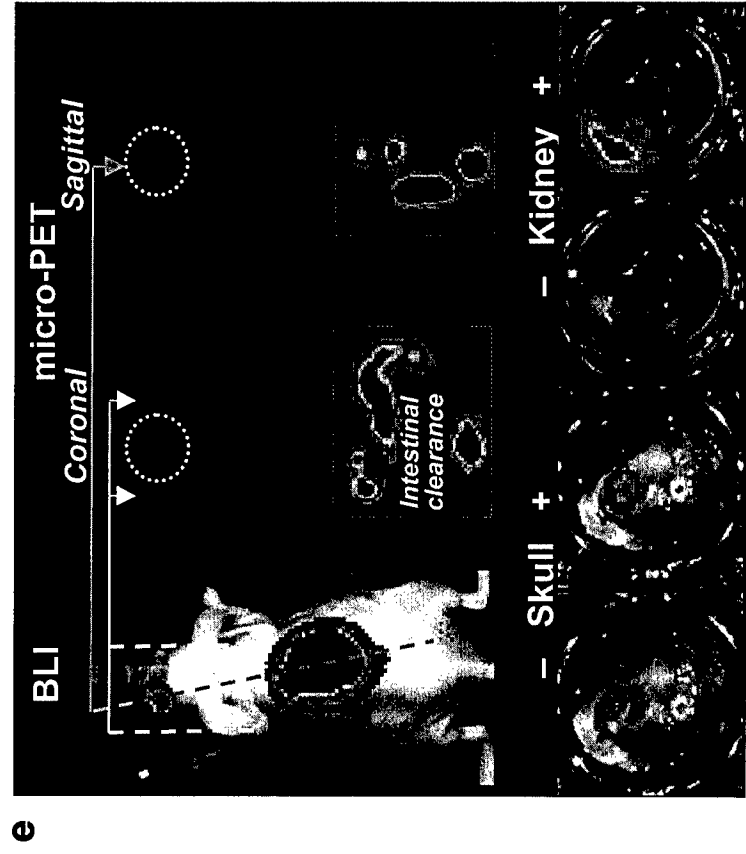
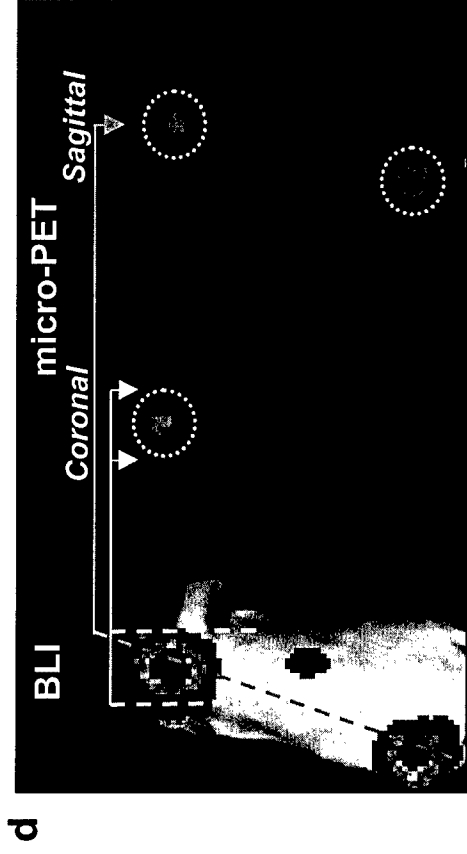
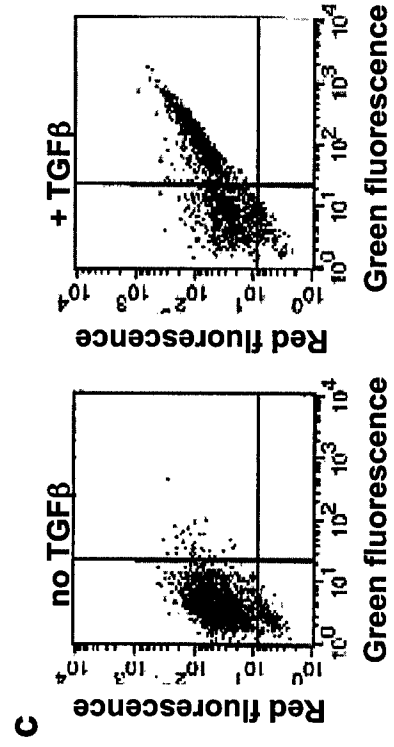
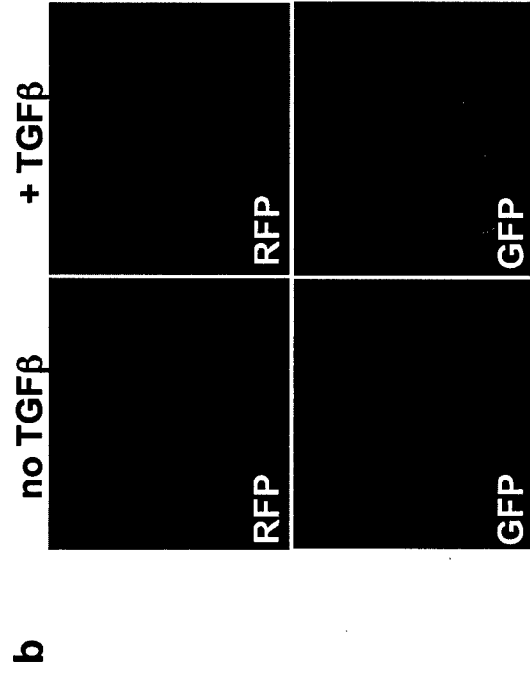
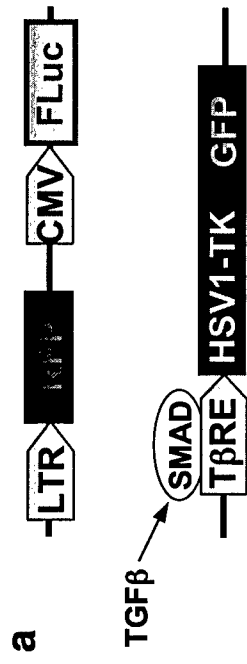
1. Mundy, G. R. Metastasis to bone: causes, consequences and therapeutic opportunities. *Nat Rev Cancer* **2**, 584-93 (2002).
2. Shi, Y. & Massague, J. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell* **113**, 685-700 (2003).
3. Derynck, R., Akhurst, R. J. & Balmain, A. TGF-beta signaling in tumor suppression and cancer progression. *Nat Genet* **29**, 117-29 (2001).
4. Siegel, P. M. & Massague, J. Cytostatic and apoptotic actions of TGF-beta in homeostasis and cancer. *Nat Rev Cancer* **3**, 807-21 (2003).
5. Derynck, R. & Zhang, Y. E. Smad-dependent and Smad-independent pathways in TGF-beta family signalling. *Nature* **425**, 577-84 (2003).
6. Dumont, N. & Arteaga, C. L. Targeting the TGF beta signaling network in human neoplasia. *Cancer Cell* **3**, 531-6 (2003).
7. Watanabe, T. et al. Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N Engl J Med* **344**, 1196-206 (2001).
8. Siegel, P. M., Shu, W., Cardiff, R. D., Muller, W. J. & Massague, J. Transforming growth factor beta signaling impairs Neu-induced mammary tumorigenesis while promoting pulmonary metastasis. *Proc Natl Acad Sci U S A* **100**, 8430-5 (2003).
9. Muraoka-Cook, R. S. et al. Conditional overexpression of active transforming growth factor beta1 in vivo accelerates metastases of transgenic mammary tumors. *Cancer Res.* **64**, 9002-9011 (2004).
10. Yin, J. J. et al. TGF-beta signaling blockade inhibits PTHrP secretion by breast cancer cells and bone metastases development. *J Clin Invest* **103**, 197-206 (1999).
11. Cailleau, R., Young, R., Olive, M. & Reeves, W. J., Jr. Breast tumor cell lines from pleural effusions. *J Natl Cancer Inst* **53**, 661-74 (1974).
12. Kang, Y. et al. A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3**, 537-49 (2003).
13. Minn, A. J. et al. Genes that mediate breast cancer metastasis to lung. *Nature* **in press** (2005).
14. Hanai, J. et al. Interaction and functional cooperation of PEBP2/CBF with Smads. Synergistic induction of the immunoglobulin germline Calpha promoter. *J Biol Chem* **274**, 31577-82 (1999).
15. Pardali, E. et al. Smad and AML Proteins Synergistically Confer Transforming Growth Factor beta1 Responsiveness to Human Germ-line IgA Genes. *J Biol Chem* **275**, 3552-3560 (2000).
16. Javed, A. et al. Impaired intranuclear trafficking of Runx2 (AML3/CBFA1) transcription factors in breast cancer cells inhibits osteolysis in vivo. *Proc Natl Acad Sci U S A.* **102**, 1454-1459 (2005).
17. Campbell, R. E. et al. A monomeric red fluorescent protein. *Proc Natl Acad Sci U S A* **99**, 7877-82 (2002).
18. Girasole, G., Passeri, G., Jilka, R. L. & Manolagas, S. C. Interleukin-11: a new cytokine critical for osteoclast development. *J Clin Invest.* **93**, 1516-1524 (1994).
19. Sotiriou, C. et al. Interleukins-6 and -11 expression in primary breast cancer and subsequent development of bone metastases. *Cancer Lett* **169**, 87-95 (2001).

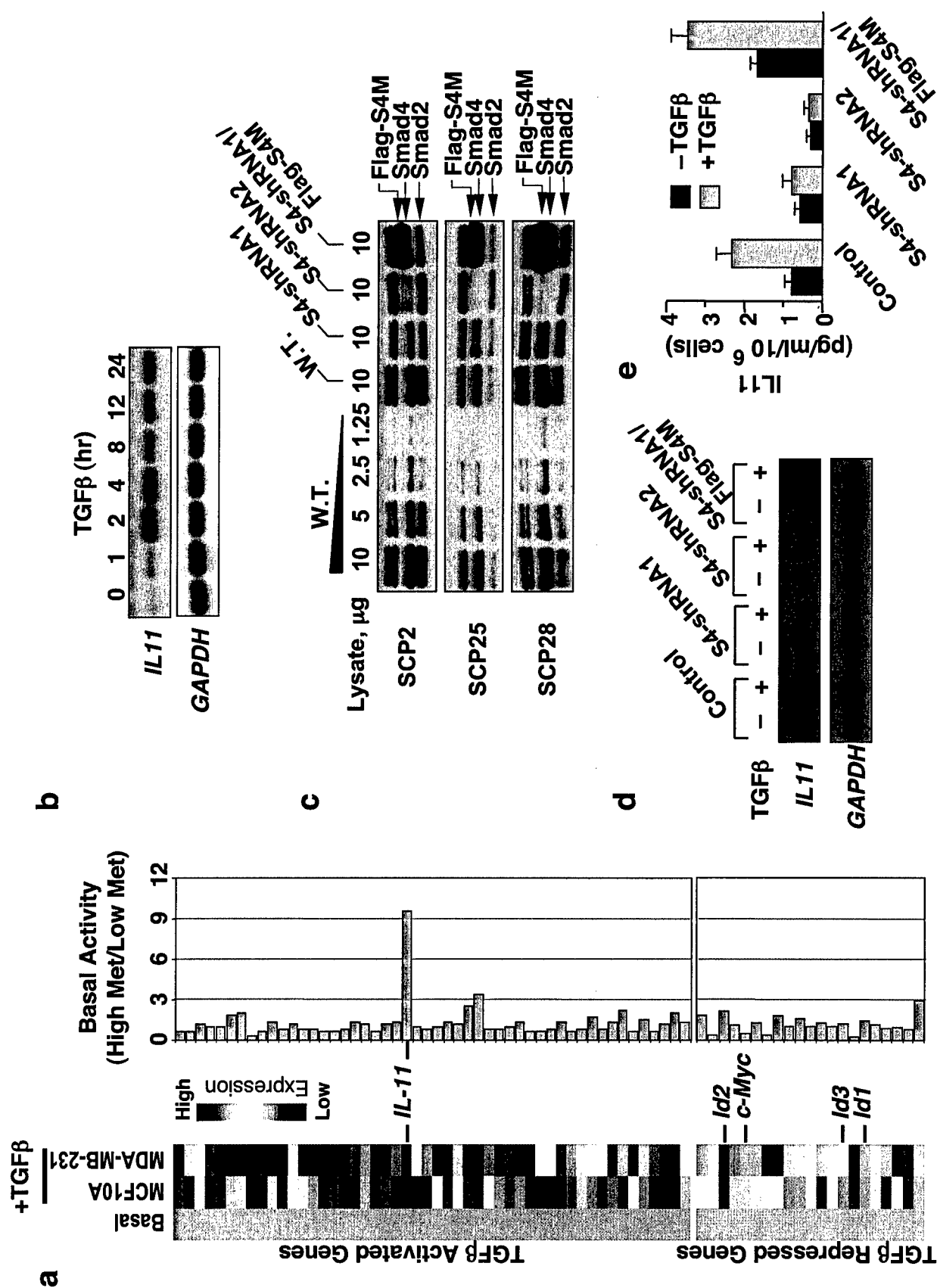
20. Morgan, H., Tumber, A. & Hill, P. A. Breast cancer cells induce osteoclast formation by stimulating host IL-11 production and downregulating granulocyte/macrophage colony-stimulating factor. *Int J Cancer*. **109**, 653-660 (2004).
21. Tang, W., Yang, L., Yang, Y. C., Leng, S. X. & Elias, J. A. Transforming growth factor-beta stimulates interleukin-11 transcription via complex activating protein-1-dependent pathways. *J Biol Chem* **273**, 5506-13 (1998).
22. Chen, C. R., Kang, Y. & Massague, J. Defective repression of c-myc in breast cancer cells: A loss at the core of the transforming growth factor beta growth arrest program. *Proc Natl Acad Sci U S A* **98**, 992-9 (2001).
23. Kang, Y., Chen, C. R. & Massague, J. A self-enabling TGFbeta response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells. *Mol Cell* **11**, 915-26 (2003).
24. Minn, A. J. et al. Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J Clin Invest* **115**, 44-55 (2005).
25. Bamba, S. et al. Regulation of IL-11 expression in intestinal myofibroblasts: role of c-Jun AP-1- and MAPK-dependent pathways. *Am J Physiol Gastrointest Liver Physiol*. **285**, G529-538 (2003).
26. Schutte, M. et al. *DPC4* gene in various tumor types. *Cancer Res*. **56**, 2527-2530 (1996).
27. Eferl, R. & Wagner, E. F. AP-1: a double-edged sword in tumorigenesis. *Nat Rev Cancer* **3**, 859-68 (2003).
28. Ponomarev, V. et al. A novel triple-modality reporter gene for whole-body fluorescent, bioluminescent, and nuclear noninvasive imaging. *Eur J Nucl Med Mol Imaging* **31**, 740-51 (2004).
29. Contag, C. H. et al. Visualizing gene expression in living mammals using a bioluminescent reporter. *Photochem Photobiol* **66**, 523-31 (1997).
30. Yingling, J. M., Blanchard, K. L. & Sawyer, J. S. Development of TGF-beta signalling inhibitors for cancer therapy. *Nat Rev Drug Discov*. **3**, 1011-1022 (2004).
31. Jakubowiak, A. et al. Inhibition of the transforming growth factor beta 1 signaling pathway by the AML1/ETO leukemia-associated fusion protein. *J Biol Chem* **275**, 40282-7 (2000).
32. Ponomarev, V. et al. Imaging TCR-dependent NFAT-mediated T-cell activation with positron emission tomography in vivo. *Neoplasia* **3**, 480-8 (2001).
33. Brummelkamp, T. R., Bernards, R. & Agami, R. Stable suppression of tumorigenicity by virus-mediated RNA interference. *Cancer Cell* **2**, 243-7 (2002).
34. Morgenstern, J. P. & Land, H. Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Res* **18**, 3587-96 (1990).
35. Chen, C. R., Kang, Y., Siegel, P. M. & Massague, J. E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression. *Cell* **110**, 19-32 (2002).
36. Serganova, I. et al. Molecular imaging of temporal dynamics and spatial heterogeneity of hypoxia-inducible factor-1 signal transduction activity in tumors in living mice. *Cancer Res* **64**, 6101-8 (2004).



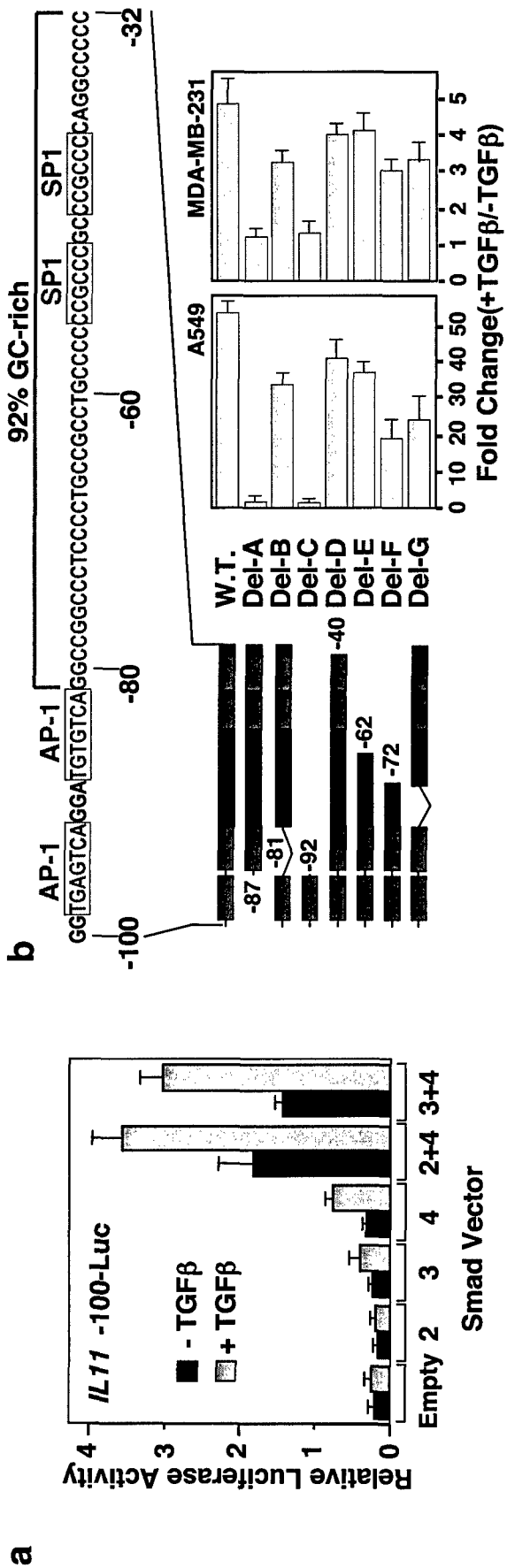
a

Case #	Phospho-Smad2	
	Tumor	Stroma
1	+++	++
2	-	-
3	+	-
4	++	++
5	+++	++
6	+++	++
7	+++	+++
8	+	-
9	+++	++
10	+++	++
11	+++	++
12	+++	++
13	-	-
14	+++	++
15	+++	++
16	+++	+++

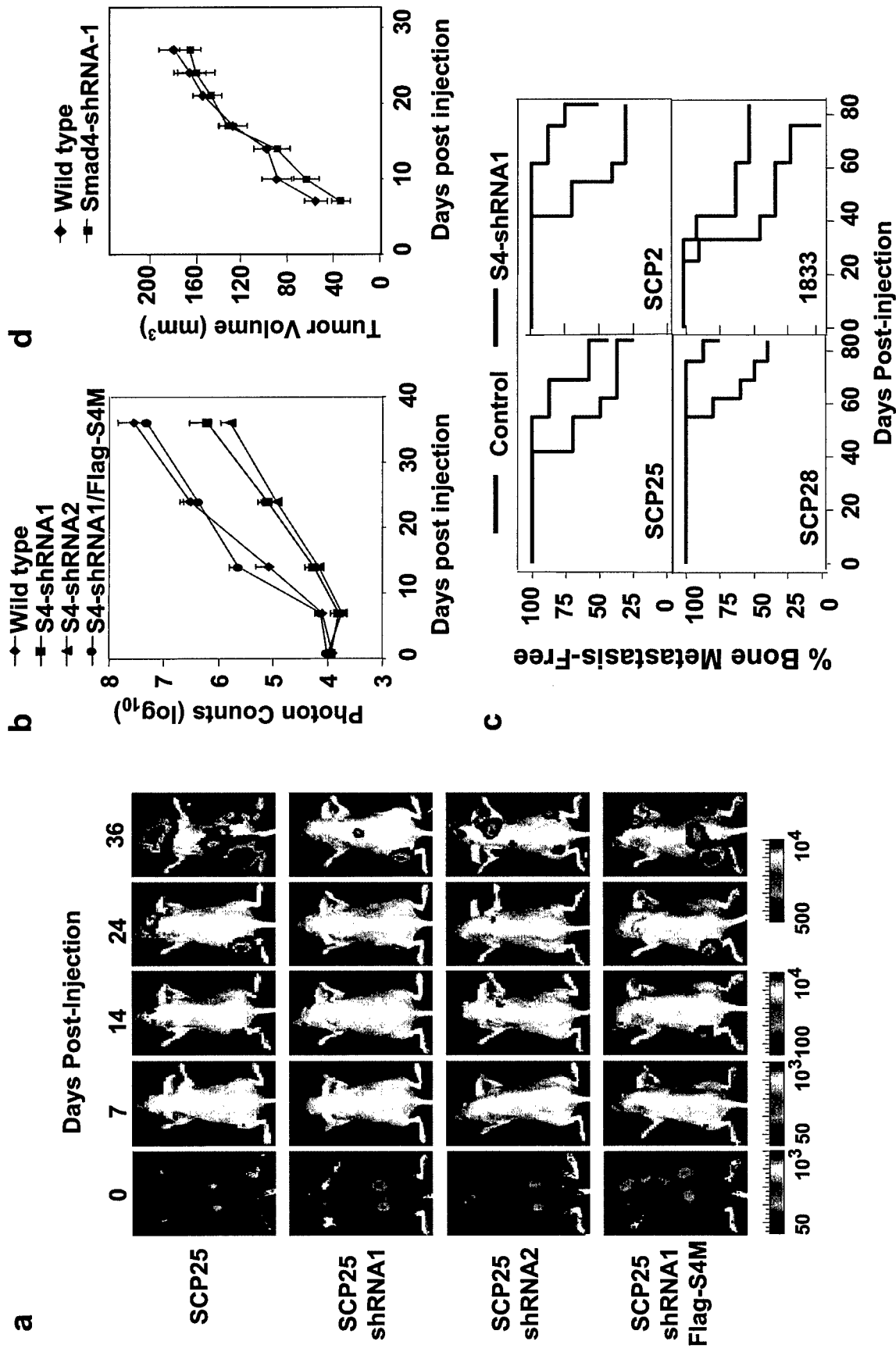




Kang et al., Figure 3



Kang et al., Figure 4



Kang et al., Figure 5

Supplementary Table 1.

Probe Set	Description	Cell Lines			
		HaCaT	MCF-10A	HPL1	MDA231
201170_s_at	Basic helix-loop-helix domain containing, class B, 2				
201329_s_at	V-ets avian erythroblastosis virus E26 oncogene homolog 2				
201389_at	Integrin, alpha 5 (fibronectin receptor, alpha polypeptide)				
201416_at	SRY (sex determining region Y)-box 4				
201466_s_at	c-jun				
201473_at	Jun B proto-oncogene				
201739_at	Serum/glucocorticoid regulated kinase				
202149_at	Enhancer of filamentation (HEF1)				
202150_s_at	Enhancer of filamentation 1 (cas-like docking; Crk-associated substrate related)				
202284_s_at	p21Cip1				
202628_s_at	plasminogen activator inhibitor type 1				
202672_s_at	activating transcription factor 3 (ATF3)				
203592_s_at	Follistatin-like 3 (secreted glycoprotein)				
204255_s_at	Vitamin D (1,25- dihydroxyvitamin D3) receptor				
204790_at	Smad7				
205330_at	Meningioma (disrupted in balanced translocation) 1, MN1				
205387_s_at	Chorionic gonadotropin, beta polypeptide				
205479_s_at	Plasminogen activator, urokinase				
205596_s_at	E3 ubiquitin ligase Smurf2				
205807_s_at	Tuftelin 1				
206277_at	Purinergic receptor P2Y, G-protein coupled, 2				
206675_s_at	Sno				
206924_at	Interleukin 11				
207147_at	Distal-less homeo box 2				
207530_s_at	p15Ink4b				
207574_s_at	Growth arrest and DNA-damage-inducible, beta				
208083_s_at	integrin, beta 6 (ITGB6)				
208322_s_at	Sialyltransferase 4A (beta-galactosidase alpha-2,3-sialyltransferase)				
209098_s_at	Jagged 1 (Alagille syndrome)				
209101_at	Connective tissue growth factor				
209193_at	Pim-1 oncogene				
209681_at	Solute carrier family 19 (thiamine transporter), member 2				
209706_at	NK homeobox (Drosophila), family 3, A				
209765_at	A disintegrin and metalloproteinase domain 19 (meltrin beta)				
210214_s_at	Bone morphogenetic protein receptor, type II (serine/threonine kinase)				
210999_s_at	Growth factor receptor-bound protein 10				
211165_x_at	EphB2				
211527_x_at	Vascular endothelial growth factor				
211981_at	Collagen, type IV, alpha 1				
212666_at	E3 ubiquitin ligase Smurf1				
213039_at	Rho-specific guanine nucleotide exchange factor p114				
216199_s_at	mitogen-activated protein kinase kinase 4				
216268_s_at	Jagged 1 (Alagille syndrome)				
217227_x_at	Immunoglobulin lambda locus				
217875_s_at	Transmembrane, prostate androgen induced RNA				
219257_s_at	Sphingosine kinase 1				
219682_s_at	T-box 3 (ulnar mammary syndrome)				
219825_at	Cytochrome P450 retinoid metabolizing protein				
221009_s_at	Angiopoietin-like 4				
221029_s_at	Wingless-type MMTV integration site family, member 5B				
201008_s_at	Thioredoxin interacting protein		D	D	D
201010_s_at	Thioredoxin interacting protein		D	D	D
201565_s_at	Id2	D	D	D	
202068_s_at	Low density lipoprotein receptor (familial hypercholesterolemia)		D	D	D
202431_s_at	c-myc	D	D	D	
202436_s_at	Cytochrome P450	D	D		D
202657_s_at	Transcriptional regulator interacting with the PHS-bromodomain 2	D	D	D	D
203973_s_at	CCAAT/enhancer binding protein (C/EBP), delta		D	D	D
204011_at	Sprouty (Drosophila) homolog 2	D	D	D	
204881_s_at	UDP-glucose ceramide glucosyltransferase	D	D	D	D
204897_at	Prostaglandin E receptor 4 (subtype EP4)		D	D	
205466_s_at	Heparan sulfate (glucosamine) 3-O-sulfotransferase 1		D	D	D
206170_at	Adrenergic, beta-2-, receptor, surface	D	D		
207826_s_at	Id3	D	D	D	D
207980_s_at	Cbp/p300-interacting transactivator, 2	D	D	D	D
208937_s_at	Id1	D	D	D	
209567_at	Homolog of yeast ribosome biogenesis regulatory protein RRS1		D	D	
210538_s_at	Baculoviral IAP repeat-containing 3	D	D	D	D
211518_s_at	Bone morphogenetic protein 4	D		D	D
218723_s_at	RGC32 protein	D	D	D	D
39402_at	Human interleukin 1-beta (IL1B) mRNA, complete cds.	D	D	D	D

genes activated by TGFb

genes repressed by TGFb

Supplementary Table 2.

Probeset	Weekly Metastatic					Highly Metastatic					Overall						
	Pancreatic	SCF3	SCF4	SCF6	SCF8	SCF2	SCF3	SCF2	SCF2	SCF2	Average	SEM	1633	2287	Average	SEM	CHARGE
201170_s.at	1420.7	906.9	1394.7	956	646.5	515.2	978	726	1031.9	726	763.48	24.79%	0.67	666.3	763.48	24.79%	0.67
201329_s.at	561.2	135.7	612.2	646.5	302.7	341.9	362.8	256.4	361.2	362.8	323.86	15.04%	0.67	275	323.86	15.04%	0.67
201369_s.at	545.6	393.4	593.4	302.7	500.675	540.2	514.6	479	682.2	514.6	629	22.70%	1.26	886.8	629	22.70%	1.26
201466_s.at	236.3	398	215.35	477.8	215.35	401.5	392.2	51.7	30.6	392.2	331	17.73%	1.04	131	331	17.73%	1.04
201473_s.at	273.2	122.4	202.7	477.8	199.075	401.5	266.3	256.7	516.5	266.7	365.08	20.18%	1.14	24.1	365.08	20.18%	1.14
201739_s.at	610.5	268.9	982	728.6	652	1389.1	1737.8	1305	1284.9	1305	1320.36	20.18%	2.03	905	1320.36	20.18%	2.03
202149_s.at	417.1	91.5	651.2	107.4	316.8	47.2	226.7	104.7	115.7	104.7	117.72	50.16%	0.37	96	117.72	50.16%	0.37
202150_s.at	499.5	101	391.3	59.3	262.925	108.7	307.7	83.2	258.9	83.2	185.32	48.97%	0.70	188.1	185.32	48.97%	0.70
202284_s.at	731.5	983	1032.3	283.4	757.55	1191.5	1057.8	781.6	808.8	781.6	1029.16	20.18%	1.38	1308.1	1029.16	20.18%	1.38
202628_s.at	222.5	475.3	1337.4	1655.1	1704.325	1624.8	1587.4	2008.6	2005.5	2008.6	1569.2	32.35%	0.92	619.7	1569.2	32.35%	0.92
202629_s.at	224.9	221.8	174	261.9	220.05	195.2	401.2	225.9	371.3	225.9	277.78	32.35%	1.26	162.5	277.78	32.35%	1.26
202592_s.at	125.9	137.1	248	241.5	188.125	82.1	215.7	151.6	134.3	151.6	150.96	30.96%	0.88	171.1	150.96	30.96%	0.88
204255_s.at	188.8	160.3	23.2	188.1	192.6	188.1	202.6	64.1	150.7	125.3	133.36	33.61%	0.69	124.1	133.36	33.61%	0.69
204790_s.at	74.1	51.3	13.3	11.9	37.575	10.3	15.4	14.1	38.2	49.5	25.5	60.76%	0.68	49.5	25.5	60.76%	0.68
205397_s.at	14.9	20.5	17.1	15.1	16.9	17.4	11	12.4	16.9	12.5	14.04	18.51%	0.83	12.5	14.04	18.51%	0.83
205478_s.at	3181.6	5112.4	4214.7	5883.2	4597.975	7512	6760.6	7607.1	4953.8	7607.1	6483.8	16.24%	1.41	5585.5	6483.8	16.24%	1.41
205596_s.at	2709.4	4276.5	3625.3	3860.8	3668	4842.3	4796.9	6063.8	2464.7	6063.8	4118.42	34.97%	1.12	2424.4	4118.42	34.97%	1.12
205807_s.at	825.3	837.2	804.6	1046.7	878.45	720	772.5	619.2	586.4	619.2	615.38	22.10%	0.70	378.8	615.38	22.10%	0.70
206277_s.at	194.3	308.9	150.8	203.1	214.525	131.4	199.3	476.2	270.7	134.5	242.42	52.60%	1.13	134.5	242.42	52.60%	1.13
206975_s.at	49.7	21.5	8.4	21.7	25.325	72.2	35.7	5.3	32.3	29.5	35	81.33%	1.38	29.5	35	81.33%	1.38
206982_s.at	206.7	360.3	78.8	163.2	207.2	164.3	247.3	200.3	1680	200.3	200.72	10.35%	0.64	1680	200.72	10.35%	0.64
207147_s.at	3.8	22.2	10.4	8.9	11.325	9.8	9	9.5	14	20.4	12.5	34.75%	1.10	20.4	12.5	34.75%	1.10
207530_s.at	11.1	14.3	9.3	3.9	9.65	5.6	4.3	11.4	13.4	4.6	7.86	48.16%	0.81	4.6	7.86	48.16%	0.81
207574_s.at	485.7	526.8	576.2	490.8	519.875	498.4	577.3	782.5	564.7	416.9	567.86	21.39%	1.09	416.9	567.86	21.39%	1.09
208083_s.at	5.6	35.4	4.2	4.6	12.45	4.1	16.6	11.3	44.1	4.9	16.2	90.58%	1.23	4.9	16.2	90.58%	1.23
208322_s.at	88.1	213.3	97.9	27.1	106.8	128.8	96.9	145.6	133.6	150.4	131.06	14.33%	1.23	150.4	131.06	14.33%	1.23
208608_s.at	255	354.5	124.7	104.5	209.675	414.9	347.4	452.8	1220.8	237.7	534.66	65.59%	2.55	237.7	534.66	65.59%	2.55
208901_s.at	1214.3	2034.5	1718.5	2016	1746.325	6924.9	7011	7901	5349.9	3587.5	5924.66	24.88%	3.40	3587.5	5924.66	24.88%	3.40
208951_s.at	300.5	585.9	169.9	586.6	523.025	324.9	474.6	412.2	491.6	360.2	452.1	17.78%	0.85	360.2	452.1	17.78%	0.85
208958_s.at	603.2	608	833.2	718.9	694.525	848.9	682.7	783.5	592.9	822.2	685.1	33.69%	0.93	822.2	685.1	33.69%	0.93
209706_s.at	274.9	340.5	311.5	225.8	288.125	333.2	391.1	481	359.5	300.4	371.04	18.69%	0.74	300.4	371.04	18.69%	0.74
209745_s.at	109.6	112.7	173.1	148.5	135.975	86.6	93.4	84.1	107.3	128.8	100.04	16.48%	0.76	128.8	100.04	16.48%	0.76
210214_s.at	1319	1352.8	1412.8	1056.3	1285.225	536.4	980.9	1132.9	1249.3	968.5	973.6	24.85%	0.83	968.5	973.6	24.85%	0.83
211165_s.at	643.5	313.3	604.6	260.9	455.575	223	329.8	394.7	910.8	666.8	535.28	41.46%	1.33	666.8	535.28	41.46%	1.33
211527_s.at	745.7	298.4	182.3	394	402.6	374.3	270.9	32.3	243.7	37.1	121.32	91.87%	0.82	37.1	121.32	91.87%	0.82
211981_s.at	407.7	14.8	129.9	172.4	181.15	22.6	27.9	34.8	415.5	345.7	396.44	16.06%	0.77	345.7	396.44	16.06%	0.77
212686_s.at	2090.6	1535	1654.2	1865.9	1836.425	2988.7	3620.6	2309.3	3166.9	3162.3	3053.56	13.97%	1.66	3162.3	3053.56	13.97%	1.66
216199_s.at	415.1	479.3	679.1	812.8	546.575	446.3	373.6	555.6	424.8	309.8	422.42	19.37%	0.77	309.8	422.42	19.37%	0.77
216288_s.at	836.6	2711.9	635.3	606	1197.45	1840.4	1250.1	2314.8	2000.8	1232.6	1697.7	25.02%	1.41	1232.6	1697.7	25.02%	1.41
217227_s.at	9.9	11.2	12.3	13.7	11.775	27.1	11.4	14.5	16	65.9	26.98	74.75%	2.29	65.9	26.98	74.75%	2.29
217875_s.at	220.5	103.8	216.1	30.3	142.675	67.8	91.5	63.8	205.3	88.5	103.34	50.47%	1.61	88.5	103.34	50.47%	1.61
219257_s.at	447.4	178.8	197.1	185	252.325	404.3	420.6	257.9	522.4	427.2	406.48	20.92%	0.63	427.2	406.48	20.92%	0.63
219682_s.at	6.2	13.1	3.4	3.8	6.625	8.7	4.3	2.8	1.9	3.3	4.2	56.86%	1.28	3.3	4.2	56.86%	1.28
219825_s.at	161	165.7	59	149.2	133.725	197.5	140.9	105.3	214.5	194.4	170.52	23.86%	1.28	194.4	170.52	23.86%	1.28
221009_s.at	169.4	176.8	73.2	91.7	127.775	307.2	112	167.3	456.9	312.3	271.14	44.75%	2.12	312.3	271.14	44.75%	2.12
221029_s.at	395.7	241.5	229.9	123.3	247.6	249.8	460.1	330.8	340.1	272.4	330.64	22.13%	1.34	272.4	330.64	22.13%	1.34
39402_s.at	138.2	138.8	78.2	64.5	109.925	347.9	277.1	156.1	165.4	637.9	316.88	55.45%	2.88	637.9	316.88	55.45%	2.88
201008_s.at	2771.6	2050.9	2437.8	1951.8	2303.025	1946	1398.3	1576.5	2838.3	1398.9	1831.6	29.57%	0.80	1398.9	1831.6	29.57%	0.80
201010_s.at	3876.1	3526.4	3967.3	3416.8	3721.65	4307.2	2951.2	3543.7	4319	2788.9	3592	18.08%	0.89	2788.9	3592	18.08%	0.89
201565_s.at	8	142.9	9.8	24.2	46.225	42.6	1.9	47.1	31.7	41.8	41.02	62.82%	1.15	41.8	41.02	62.82%	1.15
202088_s.at	4292.9	5196.8	3204.2	3882.5	4146.8	4449.7	5431.1	4100.7	4882.7	4808.3	4750.5	9.53%	1.41	4808.3	4750.5	9.53%	1.41
202431_s.at	3564	3335.9	4865.7	6476.5	4590.525	2232.5	1943.7	2066.9	3162.4	2609.1	2388.92	18.51%	0.27	2609.1	2388.92	18.51%	0.27
202495_s.at	1982.9	1625.9	948.7	1469.1	1695.65	1385.3	1211.1	1606	999.3	873.6	1253.28	26.19%	1.20	873.6	1253.28	26.19%	1.20
202657_s.at	1829.3	627.5	74.3	1469.1	1012.05	1164.3	1196	1168.9	1259.6	1259.6	1212.46	3.16%	1.01	1259.6	1212.46	3.16%	1.01
20301_s.at	897.3	1293.5	75	1353	1069.05	1024.2	898.2	630.7	895.6	2451.6	1079.7	83.43%	1.27	2451.6	1079.7	83.43%	1.27
204881_s.at	4076	2167.3	3957.3	3818.3	3478.225	2495.5	3095	3497	4397	4538	3697.3	20.18%	1.56	4538	3697.3	20.18%	1.56
204897_s.at	142.1	215.4	154.7	287.3	200.125	249.5	428.7	258.8	276.3	134.2	311.7	38.95%	1.30	134.2	311.7	38.95%	1.30
205466_s.at	298.3	348.5	535.2	302.4	370.775	452	301.3	383.8	404.4	359.4	381.18	13.10%	1.03	359.4	381.18	13.10%	1.03
205617_s.at	575.3	787.6	677.2	744.8	696.225	1280.2	1372.3	1630.8	893.5	1096	1254.56	19.90%	1.80	1096	1254.56	19.90%	1.80
207890_s.at	199.8	384.9	592.3	476.2	493.175	215.2	212.8	135.6	146.8	259.2	193.92	23.85%	0.39	259.2	193.92	23.85%	0.39
208337_s.at	810.1	384.9	592.3	476.2	493.175	671.8	721.7	483.1	848.9	827.5	710.62	18.48%	1.26	827.5	710.62	18.48%	1.26
208567_s.at	820.9	795.8	778.8	753.8	787.325	1861.5	1077.8	1103.4	869.4	2198.1	1419.98	36.41%	0.49	2198.1	1419.98	36.41%	0.49
210538_s.at	374	141.2	239.9	172.8	737.325	944.6	961.6	1129.3	788.1	475	855.74	25.95%	1.09	475	855.74	25.95%	1.09
211518_s.at	3528.8	2054	3215.2	1152.4	2487.6	1063.7	326.7	135.9	348.3	608	496.12	64.56%	2.14	608	496.12	64.56	

Title: An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and proliferative response to androgen

Authors and Affiliations: Ashley S. Doane¹, Michael Danso², Priti Lal¹, Monica Donaton¹, Liying Zhang¹, Clifford Hudis², and William L. Gerald¹

Departments of ¹Pathology and ²Medicine, Memorial Sloan-Kettering Cancer Center, 1275 York Ave, New York, NY, United States, 10021.

Reprint requests to William L. Gerald, M.D., Ph.D. geraldw@mskcc.org

Running title: Unique estrogen receptor-negative breast cancer subtype

Key Words: breast cancer, estrogen receptor-negative, androgen responsive, gene expression

Abstract

Little is known of the underlying biology of estrogen receptor-negative, progesterone receptor-negative [ER(-)/PR(-)]breast cancer (BC) and few targeted therapies are available. Clinical heterogeneity of ER(-)/PR(-) tumors suggests that molecular characterization may provide insight into their biology, reveal distinct subsets and identify new therapeutic targets. We performed genome-wide expression analysis of 99 primary BC samples and 8 BC cell lines and identified a subset of ER(-)/PR(-) tumors with expression of genes known to be either direct targets of ER, responsive to estrogen, or differentially expressed in ER(+) BC. Differentially expressed genes included SPDEF, FOXA1, XBP1, CYB5, TFF3, NAT1, APOD, ALCAM and AR ($p < 0.001$). A classification model based on the expression signature of this tumor class identified molecularly similar breast cancers in an independent human breast cancer data set and among breast cancer cell lines (MDA-MB-453). This cell line demonstrated a proliferative response to androgen in an androgen receptor dependent and estrogen receptor independent manner. In addition the androgen dependent transcriptional program of MDA-MB-453 significantly overlapped the molecular signature of the unique ER(-)/PR(-) subclass of human tumors. This subset of breast cancers, characterized by a hormonally related transcriptional program and proliferative response to androgen, suggest the potential for therapeutic strategies targeting the androgen signaling pathway.

Introduction

Breast cancer remains a major public health concern in the United States and is the second highest cause of cancer death in women. It is estimated that in 2005 over 200,000 women will develop breast cancer and 40,400 will die of their disease (1). The estrogen receptor (ER) regulates growth and differentiation of the normal mammary gland and is important in the development and progression of about 70% of breast cancer. Like other steroid-hormone receptors, the ER mediates its downstream effects by direct transcriptional regulation of target genes. On ligand binding, the receptor dissociates from its cytoplasmic chaperones, translocates to the nucleus, binds to specific DNA sequences called estrogen-response elements (ERE) and initiates gene transcription (2). Associated co-regulatory proteins either activate or repress ER transcriptional activity (3). In recent years alternative ER signaling via direct association with and activation of many signal transduction pathways has been described (4, 5). For several decades, targeting the ER has been the cornerstone in treatment for ER-positive [ER(+)] breast cancer. Estrogen deprivation therapy may be achieved by oophorectomy, selective estrogen receptor modulators such as tamoxifen, and more recently by the use of third generation aromatase inhibitors (6) and direct estrogen receptor antagonists (7-11).

ER-negative, progesterone receptor-negative [ER(-)/PR(-)] breast cancer represents approximately 25 to 30% of all breast cancers and generally has a more aggressive clinical course. In contrast to ER(+) breast cancer, patients with ER(-)/PR(-) tumors derive little or no benefit from anti-estrogen therapy (10) and targeted therapies remain elusive (12). One notable exception has been the successful use of antibodies

targeting the tyrosine kinase receptor HER-2-neu (ERBB2) (12). Although ERBB2 can be over expressed in both ER(+) and ER(-) breast cancer, it tends to be disproportionately found in ER(-) breast cancer (13).

In addition to ER, breast cancer cells express other nuclear hormone receptors. For example the androgen receptor (AR) is expressed in 60-80% of breast cancers and implicated in breast cancer biology (14). Recent studies have reported that among postmenopausal women, high androgen levels are associated with an increased risk of developing breast cancer (15). Furthermore, androgens can induce proliferation in breast tissue, and initiate tumor formation via the AR in animal models (16). The mechanisms by which AR contribute to the initiation and progression of breast cancer and its functional relationship to the ER are unknown. It also remains to be determined if targeting the AR could extend the benefits of hormonal therapy to women with ER(-)/PR(-), AR-positive breast cancer.

Genome-wide transcript analysis using DNA microarray technology is an important and well-established new tool in the study of human disease. The technology allows the measurement of several thousands of mRNA species simultaneously. The resulting gene expression profiles can distinguish tumor classes not evident by traditional methods (17, 18). In breast cancer, DNA microarray analysis has demonstrated that ER(+) breast cancer and ER(-)/PR(-) disease have unique molecular profiles, identified several distinct molecular subclasses and been used to predict disease recurrence (19-23). Few reports specifically focus on gene expression analysis of ER(-)/PR(-) breast cancers and are limited by small sample size (24). We report the identification and characterization of a unique ER(-)/PR(-) breast cancer subset with a hormonally regulated

gene expression signature and AR-dependent, androgen induced cell growth in culture. This represents a clinically relevant subset of ER(-)/PR(-) breast cancer for which AR may provide a useful therapeutic target.

Methods

Samples and Gene Expression Analysis

Tissue samples were obtained from therapeutic or diagnostic procedures performed as part of routine clinical management at Memorial Sloan-Kettering Cancer Center. All research procedures using human tissue were approved by the MSKCC institutional review board. Tissues were snap frozen in liquid nitrogen and stored at -80°C. Each sample was examined histologically using hemotoxylin and eosin-stained cryostat sections and enriched for areas of interest by manual trimming of tissue blocks. Total RNA was extracted from frozen tissue by homogenization in guanidinium isothiocyanate-based buffer (Trizol; Invitrogen, Carlsbad, CA), purified using RNeasy (Qiagen, Valencia, CA) and examined for quality using denaturing agarose gel. Complementary DNA was synthesized from RNA using a T7-promoter-tagged oligo-dT primer. RNA target was synthesized from cDNA by *in vitro* transcription, and labeled with biotinylated nucleotides (Enzo Biochem, Farmingdale, NY) (25). Gene expression analysis was performed using HG-U133A oligonucleotide microarrays according to the manufacturer's instructions (Affymetrix, Santa Clara, Ca). There were 99 primary breast tumors analyzed (77 invasive ductal carcinoma, 10 invasive lobular carcinomas, 7 mixed

lobular and ductal carcinomas, 4 metaplastic carcinomas and one not specified). All ER(-)/PR(-) tumors were designated invasive ductal or invasive lobular type.

Data Analysis

Signals were quantified using Affymetrix Microarray Suite 5.0 and expression values were scaled to have a mean expression of 500 across the central 96% of values for each array. Each sample was individually characterized by both probe set intensity values and associated clinical data. A master gene table was compiled, in which specific genes represented by GenBank accession numbers were identified for each probe set (<http://www.affymetrix.com>). Annotation information corresponding to the GenBank accession number for each probe set was retrieved from the GenBank, LocusLink, Unigene, and Gene Ontology Consortium databases. All annotation information was downloaded through the Silicon Genetics Mirror server using the GeneSpider tool (GeneSpring, Silicon Genetics, Redwood City, CA).

Prior to unsupervised analyses, the gene expression measurements were filtered and normalized using the following methods. We included probe sets that varied the most across samples. Additionally, a probe set was included only if >10% of its measurements exceeded the per-chip mean of 500. For each array, probe set values were \log_2 transformed and centered to median=0. Normalization was performed so that all measurements for that array were multiplied by a scaling factor S such that the sum of the squares of the values equaled 1. Each probe set measurement was centered and normalized across samples according to the same procedure. Filtering and normalization were performed independently for each analysis.

Two-way unsupervised hierarchical clustering was performed using the software Cluster 3.0 (26) and Genespring. To cluster data, we used an uncentered standard correlation (Pearson correlation around zero) as our measure of similarity. In constructing dendograms, centroid linkage was used as the measure of proximity between clusters. Principal component analysis (PCA) was performed using Genespring. Principal components were calculated for a designated set of genes and samples, and the three principal components representing the greatest variance in expression were plotted in order to visualize samples in three dimensional gene expression space.

To identify differentially expressed genes between two groups, we used two different measures; fold change (ratio) between the normalized means of each ER(-) class and a student's t-test. For gene expression data generated from cultured cells exposed to different treatments, the data was filtered to include only probe sets with an absolute expression value greater than 200 in at least one condition and differential expression was evaluated by fold change between different conditions.

Immunohistochemistry (IHC)

Immunohistochemical detection was performed using streptavidin-biotin-peroxidase and microwave antigen retrieval methodology as described (25). Tissue blocks with multiple samples were prepared using a tissue arrayer (Beecher Instruments, Sun Prairie, WI). For each sample, three 0.6mm core sections of tissue were extracted from diagnostic areas of formalin-fixed, paraffin-embedded tissues. We defined Her2 positivity as 3+ by IHC, or 2+ by IHC with gene amplification. For ER, PR, AR, and ER β , samples were considered positive if greater than 10% of cell nuclei were

immunoreactive. Semi-quantitative analysis of ER expression was performed using whole sections obtained from the original paraffin-embedded tissue samples. Signal intensity was graded on a scale of 0-3. A final IHC score was computed by multiplying the percent of positive nuclei by the intensity.

Cell Culture

The breast cancer cell lines MDA-MB-453, MDA-MB-231, SKBR-3, HCC-1937, ZR75-1, MCF7, BT-474 and T-47D were obtained from American Type Culture Collection (<http://www.atcc.org>). Cells were maintained at 37° C in a humidified atmosphere containing 5% CO₂, in 75 cm² flasks containing Minimal Essential Medium (MEM) supplemented with 10% fetal bovine serum, 2% l-glutamine, NEAA, 1mM sodium pyruvate, 1.5g/L sodium bicarbonate, 100 I.U./ml penicillin and 100µg/ml streptomycin. Cells were passaged every 3-4 days when they reached 80% confluence, and harvested with 0.25% trypsin/EDTA.

For cell proliferation studies, cells were pelleted by centrifugation and resuspended in medium containing phenol red-free MEM supplemented with 10% charcoal-stripped fetal bovine serum (CSFBS) (Hyclone, Logan, UT), 2% l-glutamine, NEAA, 1mM sodium pyruvate, and 1.5g/L sodium bicarbonate. Cells were plated in replicates of 6 at a density of 1×10^4 cells/well in 96 well microtiter plates. 24 hours after seeding, cells were treated with various reagents and media and reagents were replenished every 3 days. Reagents used were 10 nM E2 (Sigma-Aldrich, St. Louis, MO), 0.1-10 nM R-1881 (Sigma), 10 µM flutamide (Sigma), 100 nM 4-OHT (tamoxifen) (Sigma), and 100 nM antiestrogen ICI 182780 (fulvestrant, ICI) (Tocris, Ellisville, MO).

Cell viability and proliferation were measured using the 3-(4,5 dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) colorimetric assay (American Type Culture Collection, Rockville, MD) (27) and quantified by measuring absorbance at 570nm (Victor V7 microplate reader, Perkin Elmer, Wellesly, MA).

Genome-wide expression profiling was performed for MDA-MB-453 cells in six experimental conditions that included incubation with combinations of androgen, AR antagonist, and vehicle control. The six expression time course experiments, referred to as experiments I thru VI, were performed simultaneously. Cells were grown to confluence in one 125cm² flask, trypsinized, resuspended and seeded in six 75cm² flasks at a density of 1×10^6 cells per flask. Cells were then incubated in media containing 10% FBS until 60% confluence, washed with ice cold PBS and treated with media and reagents according to the six experimental conditions. Experiment I incubated cells in media containing 10% FBS; Experiment II used charcoal stripped media supplemented with vehicle control; Experiment III used stripped media with 1nM R-1881; Experiment IV used stripped media with 1nM R-1881 and 10 μ M flutamide. For I-IV RNA was extracted after 48 hours. In experiments V and VI, cells were incubated in stripped media for 48 hours then exposed to either 1nM R-1881 (V) or vehicle control (VI) for 48 hours followed by RNA extraction.

Identification of ERE and ARE Motifs

For each probe set, GenBank accession numbers identified specific genes. 9999 bp of sequence 5' to the start of the transcription site was retrieved for all genes from the ENSEMBL database using build NCBI 34 (Version 2), updated February 2004, from the Silicon Genetics website (<http://www.silicongenetics.com/Downloads/HumanGenome9999.zip>). For

genes of interest, sequence within 1 to 5,000 bp upstream of the transcription site was analyzed for homology to the ERE consensus 5'-GGTCAnnnTGACC-3' and the ARE consensus 5'-AGAACAnnnTGTTCT-3'. We allowed for two single point discrepancies in each sequence homology analysis. For genes identified as having putative regulatory sequences, a false positive probability was estimated by observing both the frequency of the regulatory sequence upstream of all other genes, and the frequency of the regulatory sequence within a random distribution of bases. In the latter case, the percent occurrence of each base in the random distribution is set to equal the percent occurrence of each base within the sequence in question. Genes with homologous response elements were reported if the higher p-value obtained from these two observations was less than 0.0001.

Class Prediction

A prediction algorithm was developed in order to identify samples which expressed a relevant gene signature. Tissue samples were assigned to a subclass based on our unsupervised hierarchical clustering of ER(-)/PR(-) tumors. Differentially expressed genes between the two clusters (designated classes A and B) were ranked by student's t-test and those with a p-value <0.0001 were selected for use in the prediction model. The expression of each predictor gene was used to classify unknown samples using the k-nearest neighbors method (18). Based on normalized expression values, we examined 11 samples near (as measured in Euclidian Distance) the unclassified samples, and for each class, computed a p-value of the likelihood of finding the observed number of this class among the identified neighborhood members by chance, given the proportion of class membership in the training set. The class with the lowest p-value was assigned to the

unclassified sample. We specified a p-value cutoff of 0.15, so that if there was not sufficient evidence in favor of a particular class, no prediction was made. The p-value cutoff is a ratio of the p-value of the predicted class to the alternate class.

Results

Molecular heterogeneity of ER(-)/PR(-) breast cancers demonstrated by genome-wide expression analysis

In order to explore the molecular heterogeneity of breast cancers we performed genome-wide transcript profiling for 99 primary breast carcinomas using oligonucleotide microarrays. In all cases we performed immunohistochemical assessment of ER and PR to ensure the accuracy of receptor status and determine heterogeneity. Forty-one tumors were ER(-)/PR(-), 2 were ER(-)/PR(+), and 56 were ER(+). As a further evaluation of correspondence between the transcript level for ER determined by microarray and ER protein expression, we developed a semiquantitative IHC score for ER. We compared this protein expression score with the mRNA level according to the ESR1 probe set intensity, and observed a strong positive correlation (spearman rho= 0.834, $p < 0.01$) between ER protein and transcript levels. Unsupervised hierarchical clustering revealed a strong association between ER status and molecular profile as previously reported (22). However, 9 ER(-)/PR(-) breast cancers were grouped with the ER(+) tumors and 3 ER(+) samples were grouped with ER(-)/PR(-) breast cancers (fig 1). The finding of breast cancers molecularly discordant with ER status suggested heterogeneity within the major breast cancer subtypes and was further explored.

We focused our studies on ER(-)/PR(-) breast cancers and performed unsupervised hierarchical clustering limited to the 41 ER(-)/PR(-) tumors. Of the major clusters in the dendrogram, it was of particular interest that the 9 ER-discordant samples in the previous analysis were all closely correlated and contained in a single cluster with only one additional case (fig 2A). To evaluate the reproducibility of the molecular subgroups we carried out a principal component analysis and identified the three components representing the greatest variance in gene expression for the 41 ER(-)/PR(-) samples. Using the principal components to plot samples in three dimensions, these same 10 samples were distinct from the other ER(-)/PR(-) samples, demonstrating relatively robust molecular phenotypes (fig 2B). Therefore within our sample set of ER(-)/PR(-) breast cancers we detected two major molecular subdivisions: one composed of 10 samples with a molecular resemblance to ER(+) breast cancer (referred to hereafter as ER(-) class A) and another composed of the remaining 31 breast cancers (ER(-) class B).

Characterization of genes differentially expressed in ER(-)/PR(-) breast cancer subtypes

By visual inspection of two dimensional cluster diagrams it was evident that a number of gene clusters corresponding to differential expression in ER(-) class A relative to other ER(-)/PR(-) breast cancers are associated with ER(+) tumors (figure 1). These initial observations suggested that ER(-) class A tumors expressed a molecular signature common to ER(+) breast cancers and warranted further investigation. We first identified 202 genes markedly differentially expressed according to ER status (at least three fold difference between the means of ER(+) and ER(-)/PR(-) cases and a student's

t-test $p < 0.0001$). One hundred forty five genes were over expressed in ER(+) breast cancers and 57 were over expressed in ER(-)/PR(-) tumors (supplementary tables 2a and 2b). Not surprisingly many of the differentially expressed genes have been identified in previous similar analyses (21, 22). We then identified 142 genes significantly differentially expressed between ER(-)/PR(-) class A and class B samples. Ninety-six genes were over expressed and 46 genes were under expressed relative to class A. Of the 96 genes differentially over expressed in class A, 12 have been reported as experimentally valid direct targets of the ER (28), 12 were responsive to estrogen in previous genome wide molecular studies (29) and 24 were differentially over expressed in ER(+) tumors compared to all ER(-)/PR(-) tumors in our data (table 1). In addition, we searched 5Kb of DNA sequence 5' upstream of the transcription start site and found that 24 of the genes over expressed in class A had promoter regions containing at least one putative ERE, and 12 had promoter regions with at least one putative ARE. Among the 46 genes under expressed in ER(-)class A, 3 genes have been identified as experimental targets of the ER (28), 4 genes had promoter regions containing at least one putative ERE, and 2 genes had promoter regions containing at least one putative ARE. In addition, 5 genes were differentially over expressed among all ER(-)/PR(-) tumors compared to ER(+) tumors in our data (table 1). These observations suggested that ER(-) class A samples more closely resembled an ER(+) breast cancer molecular phenotype due to expression of many genes believed to be hormonally regulated based on data from several lines of investigation.

In order to further evaluate this finding we obtained unigene id numbers for 386 estrogen responsive genes identified in a previously published genome wide expression

analysis using an experimental platform different from that used in our study (29), and identified 508 corresponding affymetrix probe sets using the Netaffyx batch query tool (www.affymetrix.com). Supervised cluster analysis limited to this set of genes tended to group ER(-) class A samples between ER(+) samples and the remaining ER(-)/PR(-) tumors (supplemental figure 1). This was also true for the ER(-) class A cell line described below. This provides further evidence that ER(-) class A breast cancers were characterized by expression of estrogen associated gene profiles that are similar to those of ER(+) tumors.

Immunohistochemical analysis of gene transcript differences between ER(-)/PR(-) breast cancer subtypes

In order to further evaluate and validate the molecular differences identified by the genome-wide expression analysis using alternative techniques, we performed IHC for several genes differentially expressed between ER(-) class A and ER(-) class B. A significant proportion of ER(-) class A samples were immunoreactive for the AR and FOXA1 compared to ER(-) class B samples ($p=0.045$ and $p=0.013$ respectively, Fisher's exact test) in concordance with the transcript levels. Protein expression of ALCAM and SPDEF were analyzed on a continuous scale using an IHC score (percentage of cells staining times intensity). There was a significant correlation between protein and transcript expression for ALCAM (spearman $\rho=0.55$, $p=0.0002$) and significant differential protein expression between ER(-) class A compared to ER(-) class B samples ($p=0.023$, Mann-Whitney test). Nuclear expression of SPDEF was significantly greater

for ER(-) class A samples compared to ER(-) class B samples ($p < 0.0001$, Mann-Whitney test) (figure 3).

We also evaluated the breast cancer marker ERBB2 using IHC and FISH. The proportion of ERBB2 positive samples for ER(-) class A and ER(-) class B were 0.30 and 0.15 respectively, and was in good agreement with the ERBB2 transcript levels. IHC was also used to evaluate the expression of ER β . Luminal epithelial cells of normal breast expressed moderate levels of ER β , however there was little to no ER β protein expression detected in the ER(-)/PR(-) samples.

Because several genes differentially expressed in ER(-) class A were identifiable at the protein level by IHC in FFPE tissue sections, it may be feasible to develop a combination of IHC markers for routine clinical identification of ER(-) class A breast cancers. A combination of SPDEF and ALCAM was estimated to predict ER(-) class A with a sensitivity approaching 100% (95% C.I. 69 to 100%), and a specificity of 94% (95% C.I. 79 to 99%). It is important to note that this analysis is limited by sample size and the development of an IHC assay for routine clinical assessment deserves further study.

Class prediction and independent evaluation of ER(-) breast cancer subsets

In order to determine if the ER(-) class A subclass was a reproducible finding and identify appropriate breast cancer cell lines for further study, we developed a k-nearest neighbor classification model using 179 genes that were differentially expressed (p -value < 0.0001) between ER(-) class A and all other ER(-)/PR(-) tumors. We applied this classification method to an independent, publicly available breast cancer gene expression

data set that used the same analytical platform (30). A similar proportion of ER(-)/PR(-) samples was classified as ER(-) class A in this independent data as in our samples (32% of ER(-)/PR(-) tumors vs. 24%). Further analysis confirmed that a number of genes differentially expressed in the comparison of ER(-) class A and ER(-) class B in our original data were also differentially expressed in the independent predicted subsets. AR, CYB5, XBP1, FOXA1 and SPDEF, as well as the androgen responsive genes APOD (31) and PIP (32), were among the top 50 significantly over expressed genes in the predicted class A ($p < 1e-10$). It is interesting to note that ERBB2 and FGFR4 were also among the top ranked genes, highly over expressed in the predicted ER(-) class A ($p = 1.6e-14$ and $p = 1.4e-08$ respectively) of the independent data. Although ERBB2 was preferentially expressed in ER(-) class A in our original data ($p = 0.00125$), the highly significant difference in absolute expression in the larger predicted ER(-) class A more strongly suggested that ERBB2 may be an important factor in the molecular phenotype.

Not only was the ER(-) class A clearly distinguishable in the independent data by supervised analysis, but unsupervised approaches indicated that these classes represent a primary distinction among ER(-)/PR(-) tumors. An unsupervised hierarchical clustering of the 77 ER(-)/PR(-) tumors yielded primary groups of samples which corresponded very closely to the class prediction assignments by the predictive model (fig 4). This provided further evidence that the ER(-) class A and B distinction was reproducible and intrinsic to the primary molecular substructure of ER(-)/PR(-) tumors.

We then used the prediction model to evaluate breast cancer cell lines in order to identify ER(-)/PR(-) cell lines corresponding to the ER(-) class A molecular phenotype. Expression profiles were generated for the ER(-)/PR(-) cell lines MDA-MB-231, MDA-

MB-453, HCC-1937, and SKBR-3. These cell lines have been described to represent important distinctions within the spectrum of ER(-)/PR(-) disease (33). Our classification model identified the cell line MDA-MB-453 as ER(-) class A (p value ratio = 5.75e-06), and the remaining ER(-)/PR(-) cell lines as ER(-) class B. We have therefore used MDA-MB-453 as an *in vitro* model representing the ER(-) class A molecular phenotype.

The ER(-) class A cell line MDA-MB-453 shows a proliferative response to androgen that is AR-dependent and ER-independent.

The identification of ER(-)/PR(-) breast tumors characterized by expression profiles including estrogen regulated genes suggested an ER-independent mechanism for activation of hormonally responsive transcription that contributed to tumor growth and survival. In order to define the mechanism for regulation of this profile we first sought to determine whether low levels of active ER, below the limit of detection in clinical assays, might be contributing to growth of ER(-) class A tumors. In our group A model cell line MDA-MB-453, ER transcript levels were very low with an absolute expression of 38.0 and Affymetrix MAS 5.0 call of Absent. Incubation with 100nM E2 had no effect on cell culture growth compared to vehicle control. Accordingly, incubation with either the pure anti-estrogen ICI or tamoxifen, alone or in combination with 100nM E2, had no effect on overall cell viability compared to vehicle control (fig 5A). This is in contrast to the ER(+) cell line MCF-7 that was markedly growth stimulated by administration of 100nM E2, and this effect was abrogated by the addition of the pure anti-estrogen ICI (data not shown). These results suggested the ER was not playing an active role in the ER(-) class A cell line growth and survival.

Because there is the potential for functional overlap of transcriptional regulation by steroid hormone receptors we reasoned that other nuclear receptors might play a role in ER(-) class A breast cancers. We examined the expression of many known nuclear hormone receptors, including ER- β , ESRRA, PR, AR, RARA, RXRA, GCHR, PARP, and VDR and found that the AR was the only one strongly differentially over expressed in ER(-) class A. The AR has been implicated in the pathogenesis of breast cancer (16), and it is known to activate a number of estrogen responsive genes (34). Incubation with the synthetic non-metabolizable androgen R-1881 at concentrations between 0.1nM and 10nM stimulated growth in MDA-MB-453. This proliferative effect was abrogated by the addition of the AR antagonist flutamide, confirming that the response was AR dependent (fig 5B). Again we determined that the effects of androgen were not dependent on the ER as MDA-MB-453 cells treated with androgens in combination with the antiestrogens tamoxifen or ICI had minimal effect on the androgen-induced proliferation (fig 5C). These observations indicate that AR signaling is intact in ER(-) class A breast cancer cell lines and that cell growth and survival are responsive to androgen in an AR-dependent, ER-independent manner.

The ER(-) class A molecular phenotype is androgen dependent

Because the ER(-) class A cell line MDA-MB-453 demonstrated a proliferative response to androgen we set out to determine whether this was associated with the transcriptional program characteristic of ER(-) class A breast cancers. We monitored gene expression changes after administration of androgens, androgen antagonists, or vehicle control to the ER(-) class A cell line MDA-MB-453 under a variety of growth

conditions (see Methods). The results for the various experiments were concordant but the most pronounced differences in gene expression were observed between those cells first incubated in steroid deprived conditions for 48 hours, and then treated with either R-1881 or vehicle for 48 hours. After trimming to eliminate genes with very low level expression (<200 in both conditions), 497 genes were differentially expressed by at least two fold between cells exposed to R-1881 or vehicle. The androgen regulated gene SARG was upregulated by 247 fold, and has been previously shown to contain an experimentally verified, hormonally active androgen response element (35). Several other androgen responsive genes including FASN, NDRG1, and SORD each contain putative androgen response elements in their promoters (36, 37), and were upregulated after administration of R-1881. These observations provided indirect evidence that administration of R-1881 to MDA-MB-453 caused recruitment of an active AR transcription complex to highly specific AREs.

To evaluate the association between androgen responsive genes in MDA-MB-453 and the ER(-) class A molecular phenotype, we compared androgen induced gene expression changes to genes differentially expressed between ER(-) classes A and B. Of the 497 differentially expressed genes between cells treated with R-1881 or vehicle control, 22 were common to our 179 gene ER(-) class A expression signature, and this number of commonly expressed genes was higher than would be expected by chance alone ($p=3e-8$) (supplemental table 3). Therefore the genes that comprise the ER(-) class A molecular fingerprint were at least in part androgen responsive in the class A cell line.

To further explore the association between the ER(-) class A molecular phenotype and an androgen dependent transcription program, we performed principal component

analysis of the 41 ER(-)/PR(-) breast tumors using the 497 androgen responsive genes and plotted samples based on three principal components. The ER(-) class A and B samples formed distinct clusters (fig 6A). Furthermore, the same approach using the 77 ER(-)/PR(-) samples from the independent data set demonstrated clusters corresponding to our class predictions (fig 6B). These results suggested that the ER(-) class A molecular phenotype was partially recapitulated by the expression of genes regulated by androgen in ER(-) class A breast cancer cells.

We also determined whether genes induced by androgens in MDA-MB-453 corresponded to the transcriptional program activated by estrogens in ER(+) breast cancer cells and therefore could contribute to the molecular relationship between ER(-) class A and ER(+) breast cancers. Fifty of the 497 androgen responsive genes from our experiments were in common with the 386 estrogen responsive genes determined by an independent study using MCF7, T-47D, and MDA-MB-436 breast cancer cells (29). This number of common genes was much greater than would be expected by chance ($p=4e-16$) and suggested that androgen in AR positive ER(-)/PR(-) breast cancer cells can induce a transcriptional program that significantly overlaps with that induced by estrogen in ER(+) breast cancer cells.

Discussion

Clinicians have long recognized that the current classification of breast cancer based on HER2 status, histopathological grade and hormone receptor status does not sufficiently capture the clinical and biologic heterogeneity observed in practice. This has fueled efforts to develop more biologically and clinically meaningful classification based

on molecular features. We applied unsupervised and supervised analyses to gene expression profiles of primary breast cancers, and identified a subset of ER(-)/PR(-) tumors with a molecular signature that suggests an active hormonally regulated transcriptional program. Gene expression signatures were used to develop a predictive model that identified this subset among novel tissue samples and breast cancer cell lines. The breast cancer cell line MDA-MB-453 recapitulated the molecular phenotype and was used to investigate the biological basis for this subclass. Several molecules that can initiate signal transduction contributing to tumor growth and survival were over expressed in this tumor subset including AR, ERBB2 and FGFR4. We found that androgen enhanced growth of MDA-MB-453 in an ER-independent but AR-dependent manner. In addition, the ER(-) class A molecular phenotype was at least partially androgen regulated. Taken together, our findings help to define a distinctive molecular subset of ER(-)/PR(-) breast cancer with the potential for novel targeted therapeutic strategies.

The potential for molecular subclassification of breast cancers based on genome wide expression analysis has been well documented in previous studies. Applying a class discovery approach using cDNA microarrays, Perou et al (20) identified at least 5 molecular subtypes of breast cancer (termed luminal subtypes A and B, ERBB2, basal, and normal breast like). These subtypes have been repeatedly observed in independent data sets and across various high throughput platforms (38, 39). The luminal subtype A and basal groups have been the most robust in independent data analysis. This luminal subtype is primarily composed of ER(+) tumors, generally demonstrates a better prognosis and is characterized by relative over expression of estrogen related genes such

as GATA3, XBP1, FOXA1, CCND1, TFF3 and ER α . The basal class is so named because its expression pattern resembles that of the basal epithelial cell component of the breast including lack of expression of ER and related genes, and expression of cytokeratins 5/6 and 17. Hierarchical clustering of our data using the intrinsic gene list revealed that the luminal A and basal subtypes were clearly evident, while the remaining subtypes were not nearly as distinct (supplemental figure 2). In particular the ER(-) class A subtype we have described tended to be poorly correlated with any one of the five subgroups. This is similar to the findings of others and suggest that other subtypes of luminal breast cancer require refinement of their molecular definition (38, 39). The ER(-) class A samples tend to be distributed among the luminal A and other non-basal cases. This subset is most distinct when clustering is not limited to the intrinsic gene set and even more so when the analysis is limited to ER(-) breast cancers. Our data suggests that ER(-) class A breast cancers bear a much closer molecular relationship to luminal or ER(+) breast cancers than to the basal subtype despite the shared ER(-) phenotype. This observation is recapitulated in the larger independent validation set of 77 ER(-) breast cancers. The same observation in two separate breast cancer cohorts suggests that this subclass of ER(-)/PR(-) breast cancer is reproducible and distinct with important implications for the diagnosis and treatment of women with ER(-)/PR(-) breast cancer.

Our studies also suggest that the AR may play an important role in regulating the molecular events associated with ER(-) class A breast cancers. Androgenic effects on the proliferation of breast cancer cell lines are highly variable (40), an observation not particularly surprising considering the heterogeneity of AR expression in breast cancer and the complexity of AR signaling. While several breast cancer cell lines appear to be

growth inhibited by the addition of androgens, a number are growth stimulated and may be androgen dependent. 5-alpha-dihydrotestosterone (DHT) inhibits the estrogen induced proliferation of MCF-7 breast cancer cells and induces a partial G1-S phase cell cycle arrest, accompanied by an increase in cdk2-associated p21 (41). Alternatively, Lippman et al suggests that androgens stimulate cell proliferation and DNA synthesis in an AR dependent manner in some cell lines (42). In agreement with our results, previous studies have reported AR dependent androgen induced proliferation in the breast cancer cell line MDA-MB-453 (40, 43). Our data further suggest that this proliferative response is associated with a hormonally regulated transcriptional program that is common to ER(-) class A breast cancers and overlaps with ER induced transcription in ER(+) tumors. However, the overlap is incomplete, and this may reflect the fact that an integrated network of signaling pathways regulates cell proliferation. We speculate that AR may act in concert with other signal transduction pathways to contribute to the ER(-) class A molecular phenotype. For example, it is well known that receptor tyrosine kinase pathways function as modulators of nuclear hormone receptor activity (44) and in this regard it is interesting that ERBB2 is differentially expressed in ER(-) class A breast cancers. ERBB2 has been shown to stabilize AR protein levels and optimize binding of AR to promoters of androgen regulated genes in prostate cancer cells (45). In the ER(-) class A breast cancer line MDA-MB-453, blocking ERBB2 with PKI166 inhibits PI3K signaling, deactivates mTOR and decreases cell proliferation (46). Given the proliferative effect of androgen on MDA-MB-453 that we have shown, the potential for cooperative crosstalk between ERBB2 signaling and AR deserves further study. In addition, the

antiproliferative effect of antiandrogens on MDA-MB-453 provides the rational for the study of antiandrogens to treat ER(-) class A breast cancer.

It is likely that some cases of ER(-) class A breast cancer are influenced by active ERBB2 signaling. However, results of unsupervised hierarchical clustering of 99 primary tumors revealed expression of ERBB2 among several sample clusters, and suggested that the expression of ERBB2 alone does not capture the molecular phenotype of class A breast cancer. Indeed, among the class A samples in our data, only 30% were ERBB2 positive. Furthermore, SKBR-3 cells have ERBB2 gene amplification and protein over expression (33), and were identified by our predictor as class B. Not surprisingly, the ERBB2 monoclonal antibody trastuzumab inhibits the growth of SKBR-3 cells (47, 48). The ER(-) class A cell line MDA-MB-453 also overexpresses ERBB2. However, MDA-MB-453 cells are not ERBB2-amplified and are resistant to the antiproliferative effects of trastuzumab (48). The expression of ERBB2 represents a biologically and clinically important feature of breast cancer, and a molecular subtype characterized by ERBB2 over expression has been proposed (20). Our observations suggest heterogeneity within the ERBB2 molecular subtype. Indeed, ERBB2 over expression exists in estrogen responsive, ER(+) breast cancer as well as ER(-) breast cancer. Further investigation into the diversity of ERBB2 signaling among various breast cancer subtypes is required.

FGFR4 is another signaling molecule which may cooperate with AR and ERBB2 to drive tumor growth in the ER(-) class A subtype of ER(-)/PR(-) breast cancer. FGFR4 is over expressed in ER(-) class A tumors and gene amplification may exist in as many as 30% of all breast cancers (49). In MDA-MB-453 cells, FGFR4 and ERBB2 have been

shown to work in concert to activate the mTOR translational pathway and regulate cyclin D1 levels (46). Simultaneous inhibition of both pathways had a stronger antiproliferative effect than either alone. In addition, FGFR4 dependent activation of the MAPK/ERK1/2 signaling cascade can drive cell proliferation via downstream initiation of cyclin D1 transcription (50). This convergence of data suggests that further investigation into the role of FGFR4, ERBB2 and AR in ER(-) class A breast cancers is warranted and that this molecular complex may provide useful therapeutic targets for as many as 25% of ER(-)/PR(-) breast cancer patients.

Acknowledgements

We would like to thank Dr. Dennis Watson for generously providing SPDEF antibody; Yixin Wang and John A. Foekens for providing unpublished data; Adam Olshen for advice and manuscript review; Louis Vargus, Yonghong Xiao, and Lishi Chen for technical assistance; and Faye Taylor for data management. We are indebted to the Pathology and Genomics Core Facilities at MSKCC for technical support. Supported by DAMD BC010104 to WG.

References

1. Jemal A, Murray T, Ward E, et al. Cancer statistics, 2005. *CA Cancer J Clin* 2005;55(1):10-30.
2. Gruber CJ, Tschugguel W, Schneeberger C, Huber JC. Production and actions of estrogens. *N Engl J Med* 2002;346(5):340-52.
3. McKenna NJ, O'Malley BW. Minireview: nuclear receptor coactivators--an update. *Endocrinology* 2002;143(7):2461-5.
4. Losel RM, Falkenstein E, Feuring M, et al. Nongenomic steroid action: controversies, questions, and answers. *Physiol Rev* 2003;83(3):965-1016.
5. Cato AC, Nestl A, Mink S. Rapid actions of steroid receptors in cellular signaling pathways. *Sci STKE* 2002;2002(138):RE9.
6. Tobias JS. Recent advances in endocrine therapy for postmenopausal women with early breast cancer: implications for treatment and prevention. *Ann Oncol* 2004;15(12):1738-47.
7. Osborne CK, Wakeling A, Nicholson RI. Fulvestrant: an oestrogen receptor antagonist with a novel mechanism of action. *Br J Cancer* 2004;90 Suppl 1:S2-6.
8. Smith IE, Dowsett M. Aromatase inhibitors in breast cancer. *N Engl J Med* 2003;348(24):2431-42.
9. Mouridsen H, Gershanovich M, Sun Y, et al. Superior efficacy of letrozole versus tamoxifen as first-line therapy for postmenopausal women with advanced breast cancer: results of a phase III study of the International Letrozole Breast Cancer Group. *J Clin Oncol* 2001;19(10):2596-606.
10. Howell A, Cuzick J, Baum M, et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet* 2005;365(9453):60-2.
11. Goss PE, Ingle JN, Martino S, et al. A randomized trial of letrozole in postmenopausal women after five years of tamoxifen therapy for early-stage breast cancer. *N Engl J Med* 2003;349(19):1793-802.
12. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* 2001;344(11):783-92.
13. Lal P, Tan LK, Chen B. Correlation of HER-2 status with estrogen and progesterone receptors and histologic features in 3,655 invasive breast carcinomas. *Am J Clin Pathol* 2005;123(4):541-6.
14. Isola JJ. Immunohistochemical demonstration of androgen receptor in breast cancer and its relationship to other prognostic factors. *J Pathol* 1993;170(1):31-5.
15. Agoff SN, Swanson PE, Linden H, Hawes SE, Lawton TJ. Androgen receptor expression in estrogen receptor-negative breast cancer. Immunohistochemical, clinical, and prognostic associations. *Am J Clin Pathol* 2003;120(5):725-31.
16. Wong YC, Xie B. The role of androgens in mammary carcinogenesis. *Ital J Anat Embryol* 2001;106(2 Suppl 1):111-25.
17. DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14(4):457-60.

18. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286(5439):531-7.
19. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415(6871):530-6.
20. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000;406(6797):747-52.
21. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 2001;98(20):11462-7.
22. Gruvberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001;61(16):5979-84.
23. Pusztai L, Ayers M, Stec J, et al. Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. *Clin Cancer Res* 2003;9(7):2406-15.
24. Nagahata T, Onda M, Emi M, et al. Expression profiling to predict postoperative prognosis for estrogen receptor-negative breast cancers by analysis of 25,344 genes on a cDNA microarray. *Cancer Sci* 2004;95(3):218-25.
25. Holzbeierlein J, Lal P, LaTulippe E, et al. Gene expression analysis of human prostate carcinoma during hormonal therapy identifies androgen-responsive genes and mechanisms of therapy resistance. *Am J Pathol* 2004;164(1):217-27.
26. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. *Bioinformatics* 2004;20(9):1453-4.
27. Mosmann T. Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J Immunol Methods* 1983;65(1-2):55-63.
28. Tang S, Han H, Bajic VB. ERGDB: Estrogen Responsive Genes Database. *Nucleic Acids Res* 2004;32(Database issue):D533-6.
29. Cunliffe HE, Ringner M, Bilke S, et al. The gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. *Cancer Res* 2003;63(21):7158-66.
30. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;365(9460):671-9.
31. Hall RE, Aspinall JO, Horsfall DJ, et al. Expression of the androgen receptor and an androgen-responsive protein, apolipoprotein D, in human breast cancer. *Br J Cancer* 1996;74(8):1175-80.
32. Carsol JL, Gingras S, Simard J. Synergistic action of prolactin (PRL) and androgen on PRL-inducible protein gene expression in human breast cancer cells: a unique model for functional cooperation between signal transducer and activator of transcription-5 and androgen receptor. *Mol Endocrinol* 2002;16(7):1696-710.
33. Lacroix M, Leclercq G. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast Cancer Res Treat* 2004;83(3):249-89.

34. Nantermet PV, Masarachia P, Gentile MA, et al. Androgenic induction of growth and differentiation in the rodent uterus involves the modulation of estrogen-regulated genetic pathways. *Endocrinology* 2005;146(2):564-78.
35. Steketee K, Ziel-van der Made AC, van der Korput HA, Houtsmuller AB, Trapman J. A bioinformatics-based functional analysis shows that the specifically androgen-regulated gene SARG contains an active direct repeat androgen response element in the first intron. *J Mol Endocrinol* 2004;33(2):477-91.
36. Dhanasekaran SM, Dash A, Yu J, et al. Molecular profiling of human prostate tissues: insights into gene expression patterns of prostate development during puberty. *Faseb J* 2005;19(2):243-5.
37. Nelson PS, Clegg N, Arnold H, et al. The program of androgen-responsive genes in neoplastic prostate epithelium. *Proc Natl Acad Sci U S A* 2002;99(18):11890-5.
38. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003;100(14):8418-23.
39. Wilson CA, Dering J. Recent translational research: microarray expression profiling of breast cancer--beyond classification and prognostic markers? *Breast Cancer Res* 2004;6(5):192-200.
40. Birrell SN, Bentel JM, Hickey TE, et al. Androgens induce divergent proliferative responses in human breast cancer cell lines. *J Steroid Biochem Mol Biol* 1995;52(5):459-67.
41. Greeve MA, Allan RK, Harvey JM, Bentel JM. Inhibition of MCF-7 breast cancer cell proliferation by 5alpha-dihydrotestosterone; a role for p21(Cip1/Waf1). *J Mol Endocrinol* 2004;32(3):793-810.
42. Lippman M, Bolan G, Huff K. The effects of androgens and antiandrogens on hormone-responsive human breast cancer in long-term tissue culture. *Cancer Res* 1976;36(12):4610-8.
43. Hall RE, Birrell SN, Tilley WD, Sutherland RL. MDA-MB-453, an androgen-responsive human breast carcinoma cell line with high level androgen receptor expression. *Eur J Cancer* 1994;30A(4):484-90.
44. Shao D, Lazar MA. Modulating nuclear receptor function: may the phos be with you. *J Clin Invest* 1999;103(12):1617-8.
45. Mellinghoff IK, Vivanco I, Kwon A, Tran C, Wongvipat J, Sawyers CL. HER2/neu kinase-dependent modulation of androgen receptor function through effects on DNA binding and stability. *Cancer Cell* 2004;6(5):517-27.
46. Koziczak M, Hynes NE. Cooperation between fibroblast growth factor receptor-4 and ErbB2 in regulation of cyclin D1 translation. *J Biol Chem* 2004;279(48):50004-11.
47. Mayfield S, Vaughn JP, Kute TE. DNA strand breaks and cell cycle perturbation in herceptin treated breast cancer cell lines. *Breast Cancer Res Treat* 2001;70(2):123-9.
48. Yakes FM, Chinratanalab W, Ritter CA, King W, Seelig S, Arteaga CL. Herceptin-induced inhibition of phosphatidylinositol-3 kinase and Akt is required for antibody-mediated effects on p27, cyclin D1, and antitumor action. *Cancer Res* 2002;62(14):4132-41.
49. Dickson C, Spencer-Dene B, Dillon C, Fantl V. Tyrosine kinase signalling in breast cancer: fibroblast growth factors and their receptors. *Breast Cancer Res* 2000;2(3):191-6.

50. Koziczak M, Holbro T, Hynes NE. Blocking of FGFR signaling inhibits breast cancer cell proliferation through downregulation of D-type cyclins. *Oncogene* 2004;23(20):3501-8.

3

						Gene Name	p-value	Class A v B Fold Change	Common Name	Description
						214451_at	5.04E-11	44.66	TFAP2B	transcription factor AP-2 beta (activating enhancer
						209173_at	7.88E-07	25.86	AGR2	anterior gradient 2 homolog (Xenopus
						217276_x_at	2.36E-08	22.61	dJ222E13.1	kraken-like
						216623_x_at	1.14E-05	20.46	TNRC9	trinucleotide repeat containing 9
						204607_at	1.81E-09	20.06	HMGCS2	3-hydroxy-3- methylglutaryl-Coenzyme
						217284_x_at	2.56E-07	19.42	dJ222E13.1	kraken-like
						206509_at	1.02E-08	17.82	PIP	prolactin-induced protein
						214774_x_at	2.29E-06	16.7	TNRC9	trinucleotide repeat containing 9
						214243_s_at	5.46E-08	15.55	dJ222E13.1	kraken-like
						207802_at	6.88E-06	14.47	CRISP3	cysteine-rich secretory protein 3
						206463_s_at	1.37E-06	14.41	DHRS2	dehydrogenase/reductas e (SDR family) member 2
						220192_x_at	2.05E-10	13.99	SPDEF	SAM pointed domain containing ets
						209309_at	1.78E-05	13.8	AZGP1	alpha-2-glycoprotein 1, zinc
						214404_x_at	1.70E-13	12.31	SPDEF	SAM pointed domain containing ets
						215108_x_at	1.58E-05	11.79	TNRC9	trinucleotide repeat containing 9
						217528_at	7.75E-05	11.33	CLCA2	chloride channel, calcium activated, family member
						201525_at	1.80E-07	11.05	APOD	apolipoprotein D
						218922_s_at	7.43E-08	9.676	LASS4	LAG1 longevity assurance homolog 4 (S.
						213441_x_at	7.99E-11	9.494	SPDEF	SAM pointed domain containing ets
						204667_at	1.80E-12	7.782	FOXA1	forkhead box A1
						217562_at	5.78E-05	7.704	DBCCR1L	DBCCR1-like
						209813_x_at	5.71E-05	7.691	TRGV9; V2; T	T-cell receptor (V-J-C) precursor; Human T-cell

Table 1

						214079_at	4.21E-05	7.428	DHRS2	dehydrogenase/reductase (SDR family) member 2
						210576_at	6.64E-05	7.337	CYP4F8	cytochrome P450, family 4, subfamily F,
						205221_at	1.64E-06	6.974	HGD	homogentisate 1,2-dioxygenase
						213884_s_at	1.04E-09	6.785	TRIM3	tripartite motif-containing 3
						211657_at	2.95E-05	6.542		
						211237_s_at	9.51E-07	6.317	FGFR4	fibroblast growth factor receptor 4
						204623_at	2.61E-07	6.007	TFF3	trefoil factor 3 (intestinal)
						218211_s_at	5.11E-12	5.895	MLPH	melanophilin
						39763_at	2.12E-10	5.599	HPX	precursor; Human hemopexin gene, exon
						204719_at	7.17E-05	5.233	ABCA8	ATP-binding cassette, sub-family A (ABC1),
						220622_at	1.24E-05	5.222	FLJ23259	hypothetical protein FLJ23259
						210013_at	3.00E-05	5.165	HPX	hemopexin
						211110_s_at	1.46E-05	4.951	AR	androgen receptor (dihydrotestosterone
						218313_s_at	2.45E-06	4.85	GALNT7	UDP-N-acetyl-alpha-D-galactosamine:polypeptid
						204942_s_at	1.31E-07	4.721	ALDH3B2	aldehyde dehydrogenase 3 family, member B2
						203722_at	1.06E-05	4.681	ALDH4A1	aldehyde dehydrogenase 4 family, member A1
						210056_at	9.47E-09	4.538	RND1	Rho family GTPase 1
						219734_at	5.96E-10	4.512	FLJ20174	hypothetical protein FLJ20174
						221584_s_at	7.57E-06	4.344	KCNMA1	potassium large conductance calcium-
						219197_s_at	8.36E-06	4.218	SCUBE2	signal peptide, CUB domain, EGF-like 2
						204462_s_at	2.71E-05	4.158	SLC16A2	solute carrier family 16 (monocarboxylic acid
						211144_x_at	9.11E-05	4.091	TRG@	T cell receptor gamma locus
						205161_s_at	4.10E-07	4.083	PEX11A	peroxisomal biogenesis factor 11A
						215465_at	2.10E-05	4.009	ABCA12	ATP-binding cassette, sub-family A (ABC1),

Table 1

						217014_s_at	3.44E-06	3.925		Homo sapiens PAC clone RP4-604G5 from 7,
						215806_x_at	5.00E-05	3.84	TRG@	T cell receptor gamma locus
						200670_at	1.22E-08	3.671	XBP1	X-box binding protein 1
						201952_at	3.35E-06	3.568	ALCAM	activated leukocyte cell adhesion molecule
						212218_s_at	1.11E-05	3.562	FASN	fatty acid synthase
						214295_at	5.13E-05	3.491		MRNA, chromosome 1 specific transcript
						212510_at	8.81E-08	3.479	KIAA0089	KIAA0089 protein
						205306_x_at	1.73E-05	3.441	KMO	kynurenine 3-monooxygenase
						207843_x_at	4.12E-07	3.426	CYB5	cytochrome b-5
						216333_x_at	2.63E-05	3.305	TNXB	tenascin XB
						215726_s_at	1.43E-06	3.224	CYB5	cytochrome b-5
						205150_s_at	8.16E-06	3.204	KIAA0644	KIAA0644 gene product
						209366_x_at	1.17E-06	3.185	CYB5	cytochrome b-5
						219956_at	5.72E-07	3.144	GALNT6	UDP-N-acetyl-alpha-D-galactosamine:polypeptid
						219429_at	4.43E-05	3.124	FA2H	fatty acid 2-hydroxylase
						218546_at	2.19E-05	3.095	FLJ14146	hypothetical protein FLJ14146
						209522_s_at	8.33E-05	3.082	CRAT	carnitine acetyltransferase
						208284_x_at	1.74E-06	3.022	GGT1	gamma-glutamyltransferase 1
						215559_at	6.47E-05	3.013	ABCC6	ATP-binding cassette, sub-family C
						218776_s_at	4.25E-08	2.859	FLJ23375	hypothetical protein FLJ23375
						204579_at	3.21E-06	2.816	FGFR4	fibroblast growth factor receptor 4
						207131_x_at	6.49E-07	2.764	GGT1	gamma-glutamyltransferase 1
						217973_at	2.84E-06	2.74	DCXR	dicarbonyl/L-xylulose reductase
						206850_at	2.54E-05	2.71	RRP22	RAS-related on chromosome 22

Table 1

						212593_s_at	1.01E-06	2.706	PDCD4	programmed cell death 4 (neoplastic)
						205160_at	2.72E-05	2.683	PEX11A	peroxisomal biogenesis factor 11A
						203740_at	3.34E-07	2.676	MPHOSPH6	M-phase phosphoprotein 6
						211417_x_at	1.79E-07	2.591	GGT1	gamma-glutamyltransferase 1
						209919_x_at	5.16E-07	2.552	GGT1	gamma-glutamyltransferase 1
						216638_s_at	2.70E-05	2.536	PRLR	prolactin receptor
						213557_at	4.36E-05	2.492		Transcribed sequences
						212594_at	1.10E-06	2.436	PDCD4	programmed cell death 4 (neoplastic)
						201941_at	5.57E-06	2.372	CPD	carboxypeptidase D
						212736_at	8.29E-05	2.333	BC008967	hypothetical gene BC008967
						218552_at	9.75E-06	2.306	FLJ10948	hypothetical protein FLJ10948
						51158_at	9.72E-07	2.24		Clone IMAGE:4816940, mRNA
						212099_at	7.79E-05	2.189	ARHB	ras homolog gene family, member B
						200810_s_at	2.67E-06	2.187	CIRBP	cold inducible RNA binding protein
						212956_at	4.61E-06	2.166		qp61g12.x1 NCI_CGAP_Co8 Homo
						200618_at	8.97E-05	2.151	LASP1	LIM and SH3 protein 1
						211596_s_at	7.34E-05	2.146	LRIG1	leucine-rich repeats and immunoglobulin-like
						213107_at	8.17E-05	2.139		yh03e12.s1 Soares infant brain 1NIB Homo sapiens
						208872_s_at	6.53E-06	2.131	DP1	polyposis locus protein 1
						215603_x_at	3.43E-05	2.122	GGT2	gamma-glutamyltransferase 2
						219396_s_at	4.27E-05	2.116	NEIL1	nei endonuclease VIII-like 1 (E. coli)
						211621_at	9.37E-06	2.105	AR	androgen receptor (dihydrotestosterone
						215299_x_at	2.29E-05	2.019	SULT1A1; PST	Human phenol sulfotransferase (STP1)
						219543_at	8.17E-05	2.014	MAWBP	MAWD binding protein

Table 1

						201940_at	5.50E-05	2.013	CPD	carboxypeptidase D
						208873_s_at	7.34E-07	2.006	DP1	polyposis locus protein 1
						37966_at	4.40E-05	0.494	PARVB	parvin, beta
						200756_x_at	4.37E-05	0.489	CALU	calumenin
						203167_at	7.48E-05	0.487	TIMP2	tissue inhibitor of metalloproteinase 2
						219785_s_at	9.20E-05	0.485	MGC15419	MGC15419 protein
						212650_at	9.18E-05	0.478	NACSIN	NPF/calponin-like protein
						200757_s_at	2.99E-06	0.477	CALU	calumenin
						211924_s_at	8.44E-06	0.463	PLAUR	plasminogen activator, urokinase receptor
						205120_s_at	1.03E-05	0.458	SGCB	sarcoglycan, beta (43kDa dystrophin-associated
						209043_at	1.61E-05	0.457	PAPSS1	3'-phosphoadenosine 5'-phosphosulfate synthase
						218629_at	4.29E-05	0.453	SMO	smoothened homolog (Drosophila)
						200755_s_at	3.07E-05	0.443	CALU	calumenin
						209204_at	2.85E-05	0.429	LMO4	LIM domain only 4
						213003_s_at	9.43E-05	0.426		7i79f07.x1 NCI_CGAP_Ov18 Homo
						202990_at	2.27E-05	0.411	PYGL	phosphorylase, glycogen; liver (Hers disease,
						200934_at	6.04E-05	0.404	DEK	DEK oncogene (DNA binding)
						221505_at	8.10E-06	0.378	ANP32E	acidic (leucine-rich) nuclear phosphoprotein
						210074_at	5.18E-06	0.377	CTSL2	cathepsin L2
						214845_s_at	1.96E-05	0.375	CALU	calumenin
						60474_at	8.81E-05	0.371	C20orf42	chromosome 20 open reading frame 42
						202620_s_at	3.64E-05	0.37	PLOD2	procollagen-lysine, 2-oxoglutarate 5-
						202236_s_at	2.99E-06	0.363	SLC16A1	solute carrier family 16 (monocarboxylic acid
						219944_at	1.86E-05	0.346	FLJ21069	hypothetical protein FLJ21069

Table 1

						202134_s_at	3.20E-05	0.332	TAZ	transcriptional co-activator with PDZ-
						216488_s_at	2.33E-06	0.326	ATP11A	ATPase, Class VI, type 11A
						202619_s_at	6.84E-06	0.318	PLOD2	procollagen-lysine, 2-oxoglutarate 5-
						218851_s_at	8.81E-05	0.315	WDR33	WD repeat domain 33
						207675_x_at	5.89E-05	0.313	ARTN	artemin
						213256_at	1.53E-05	0.31	MGC48332	hypothetical protein MGC48332
						201564_s_at	3.22E-05	0.307	FSCN1	fascin homolog 1, actin-bundling protein
						219926_at	9.90E-05	0.306	POPDC3	popeye domain containing 3
						202784_s_at	3.61E-05	0.293	NNT	nicotinamide nucleotide transhydrogenase
						209900_s_at	5.16E-05	0.283	SLC16A1	solute carrier family 16 (monocarboxylic acid
						209834_at	2.96E-05	0.249	CHST3	carbohydrate (chondroitin 6)
						213260_at	4.56E-05	0.228	FOXC1	forkhead box C1
						208103_s_at	6.99E-05	0.227	ANP32E	acidic (leucine-rich) nuclear phosphoprotein
						204285_s_at	9.09E-09	0.222	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
						209875_s_at	7.49E-06	0.217	SPP1	secreted phosphoprotein 1 (osteopontin, bone
						202235_at	1.95E-07	0.196	SLC16A1	solute carrier family 16 (monocarboxylic acid
						204750_s_at	1.34E-05	0.144	DSC2	desmocollin 2
						204286_s_at	9.42E-07	0.141	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
						209800_at	1.03E-05	0.121	KRT16	keratin 16 (focal non-epidermolytic
						204855_at	2.69E-05	0.0816	SERPINB5	serine (or cysteine) proteinase inhibitor, clade

Table 1

Up in ER(+)
3 fold
p<0.0001

Up in ER(-)
>3 fold
p<0.0001

Putative ER(+) target Putative ER(-) target

Figure Legends

Figure 1. **Molecular heterogeneity of breast cancers.** Two way hierarchical clustering was performed with 99 primary breast cancers based on 1960 genes with the greatest variance among samples. The dendrogram represents the relationship of samples. The length of the branches represents 1- the correlation coefficient between samples. A strongly differentially expressed gene cluster is enlarged and genes associated with estrogen receptor status are labeled. Samples are arranged in columns and genes in rows. Expression levels are pseudocolored red to indicate transcript levels above the median for that gene across all samples and green below the median. Color saturation is proportional to the magnitude of expression.

Figure 2. **Molecular subclasses of ER(-)/PR(-) breast cancers.** A. Two way hierarchical clustering was performed with 41 ER(-)/PR(-) breast cancers based on 1366 genes with greatest variance among samples. Samples with a molecular similarity to ER(+) breast cancers are labeled class A and the remaining as class B. A gene cluster highly differentially expressed between the two classes is enlarged and select characterized genes labeled. B. Three dimensional plot of ER(-)/PR(-) primary breast cancers based on the three principal components representing the greatest variance in gene expression across the 41 ER(-)/PR(-) samples identified by analysis of all 22,283 U133A probe sets.

Figure 3. **Immunohistochemical evaluation of differentially expressed genes.** Representative photo micrographs of immunohistochemistry studies for ALCAM in an ER(-) class A breast tumor (A) and an ER(-) class B breast tumor (B). SPDEF in a class A breast tumor (C) and class B breast tumor (D). AR in a class A breast tumor (E) and a class B breast tumor (F).

Figure 4. **Reproducibility of ER(-) breast cancer subclasses.** Two way hierarchical clustering was performed with 77 ER(-) breast tumors from an independent data set using 1262 genes with greatest variance across samples. The resulting dendrogram revealed a tendency to group samples according to our class prediction assignments. A strongly differentially expressed gene cluster is enlarged and genes associated with ER status and class A are labeled.

Figure 5. **ER(-) class A breast cancer cells proliferate in response to androgen in an AR dependent and ER independent manner.** MDA-MB-453 cells were treated with reagents as indicated and cell proliferation measured. All experiments were performed in triplicate. A. Incubation with E2, the antiestrogens tam and ICI with or without E2, and vehicle control. B. Incubation with the androgen R-1881, R-1881 with the AR antagonist flutamide, flutamide alone, and vehicle control. C. Incubation with R-1881, R-1881 with tam, R-1881 with ICI, and vehicle control.

Figure 6. **Molecular subclasses of ER(-) breast cancer based on androgen responsive genes.** Three dimensional plot of the three principal components with the greatest variance across 41 ER(-)/PR(-) primary breast tumors using 497 genes responsive to androgen in the class A cell line MDA-MB-453. B. Three dimensional plot of the three principal components with the greatest variance among 77 ER(-) breast tumors from an independent data set using the 497 androgen responsive genes. Samples are colored according to class prediction assignments.

Supplementary Figure 1. **Molecular subclasses of breast cancer based on genes responsive to estrogen.** A two way hierarchical clustering dendrogram of 99 primary breast cancers was performed using 387 estrogen responsive genes described in reference 29. Expression levels of the estrogen receptor RNA and selected genes responsive to estrogen and over expressed in class A are depicted.

Supplementary Figure 2. Two way hierarchical clustering of 99 primary breast cancers limited to genes of the intrinsic gene list of reference 20. Representative clusters of genes corresponding to subtypes described are: Luminal A gene cluster (1), ERBB2 gene cluster (2), Normal Breast Like gene cluster (3), Basal gene cluster (4), and Luminal B gene cluster (5).

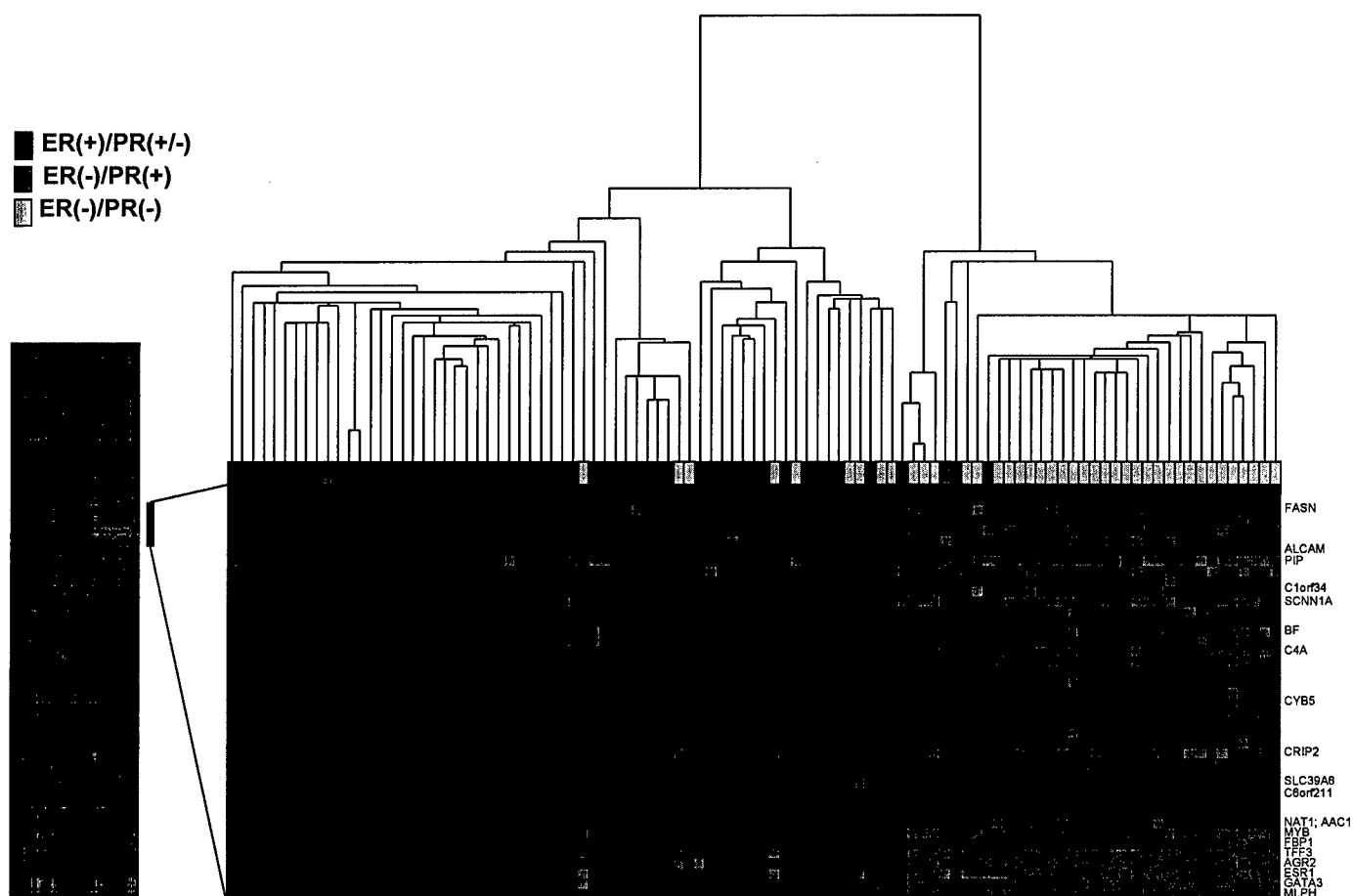
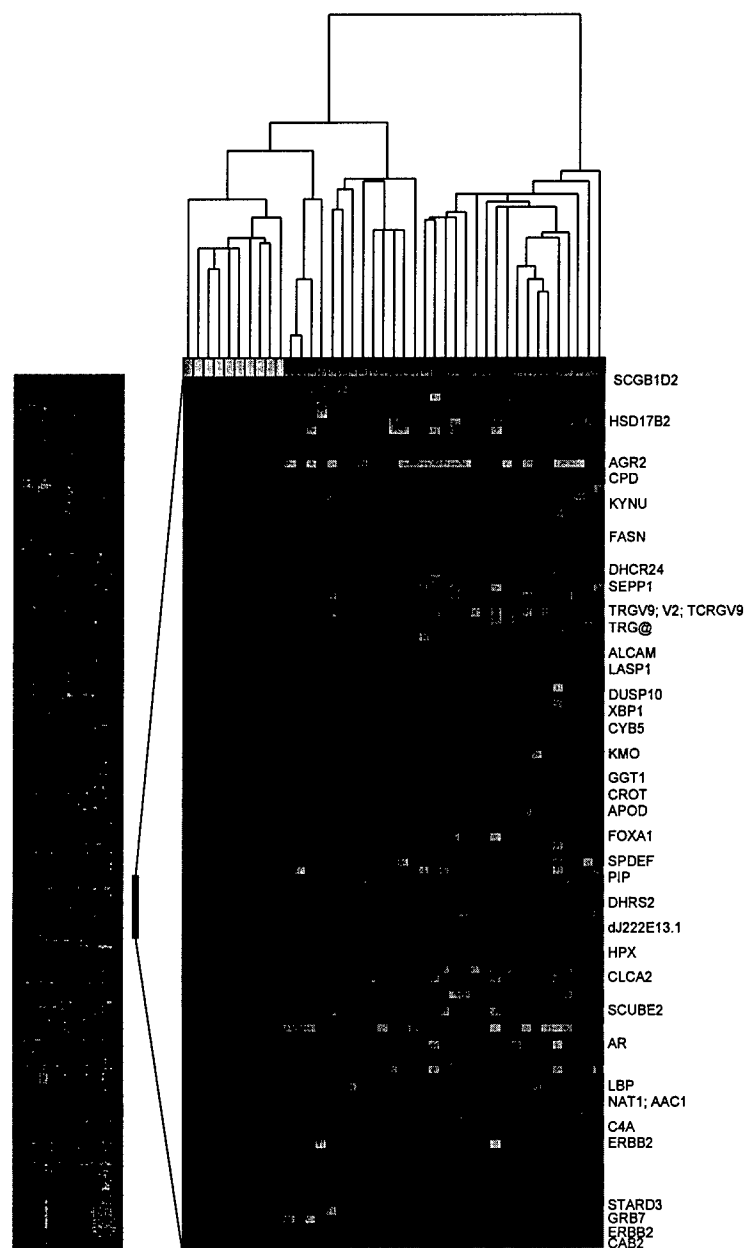
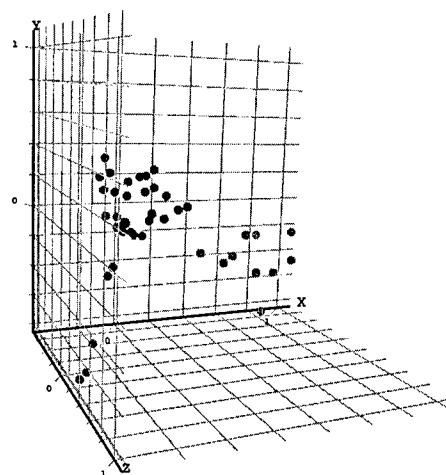
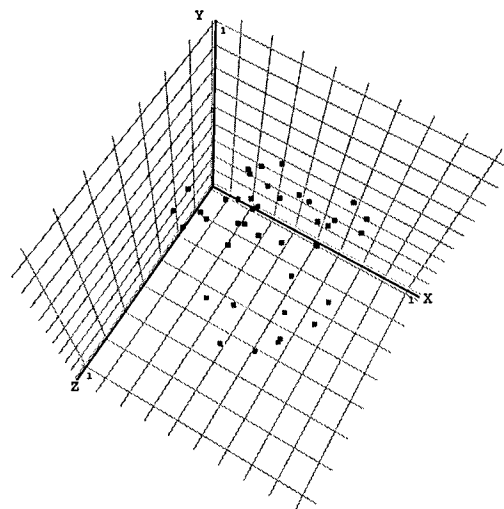


Figure 1.



A

■ Class A
■ Class B



X-axis: PCA component 1 (12.55% variance)
Y-axis: PCA component 2 (10.38% variance)
Z-axis: PCA component 3 (7.635% variance)

B

Figure 2.

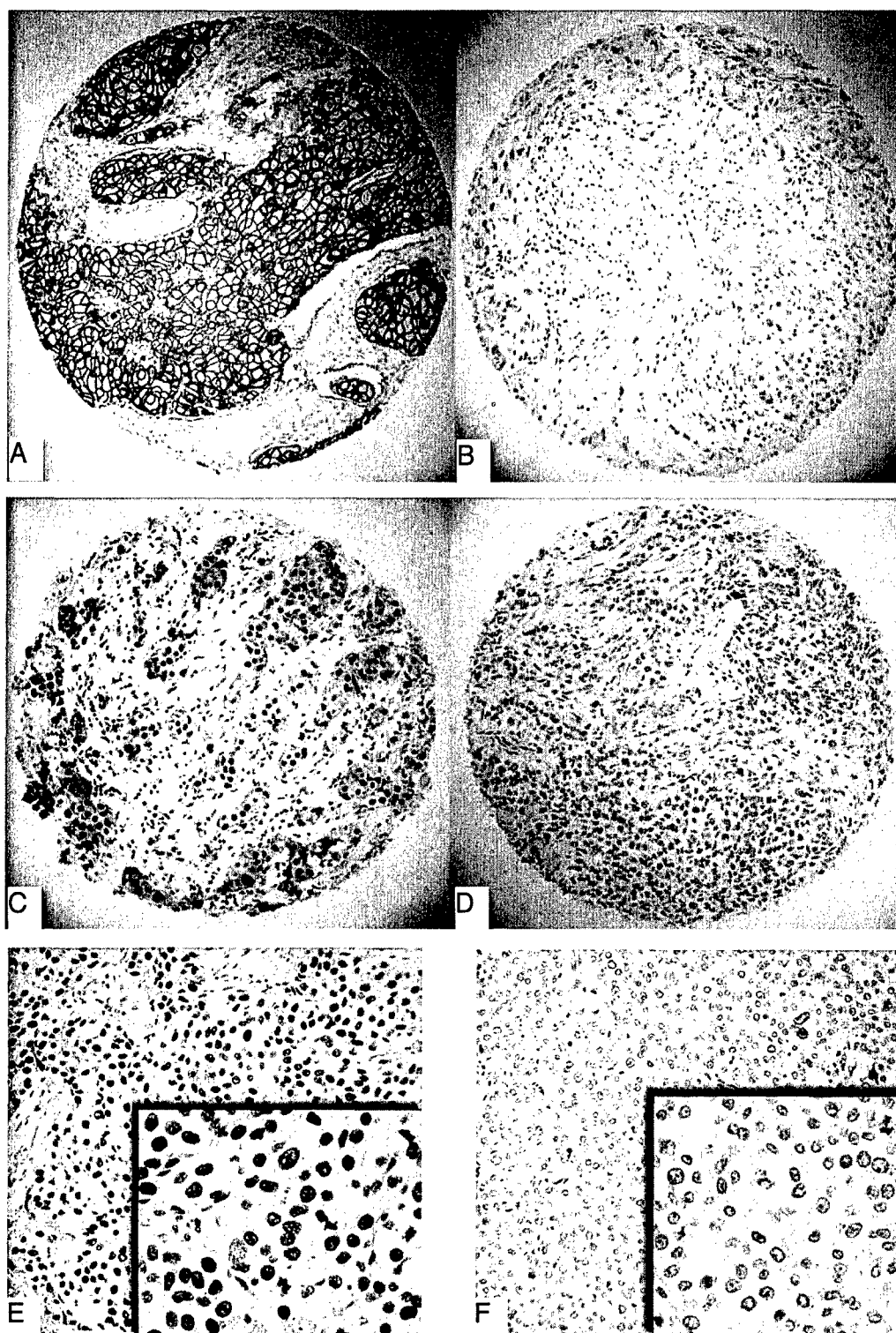





Figure 3.

 Predicted Class A
 Predicted Class B
 Unclassified

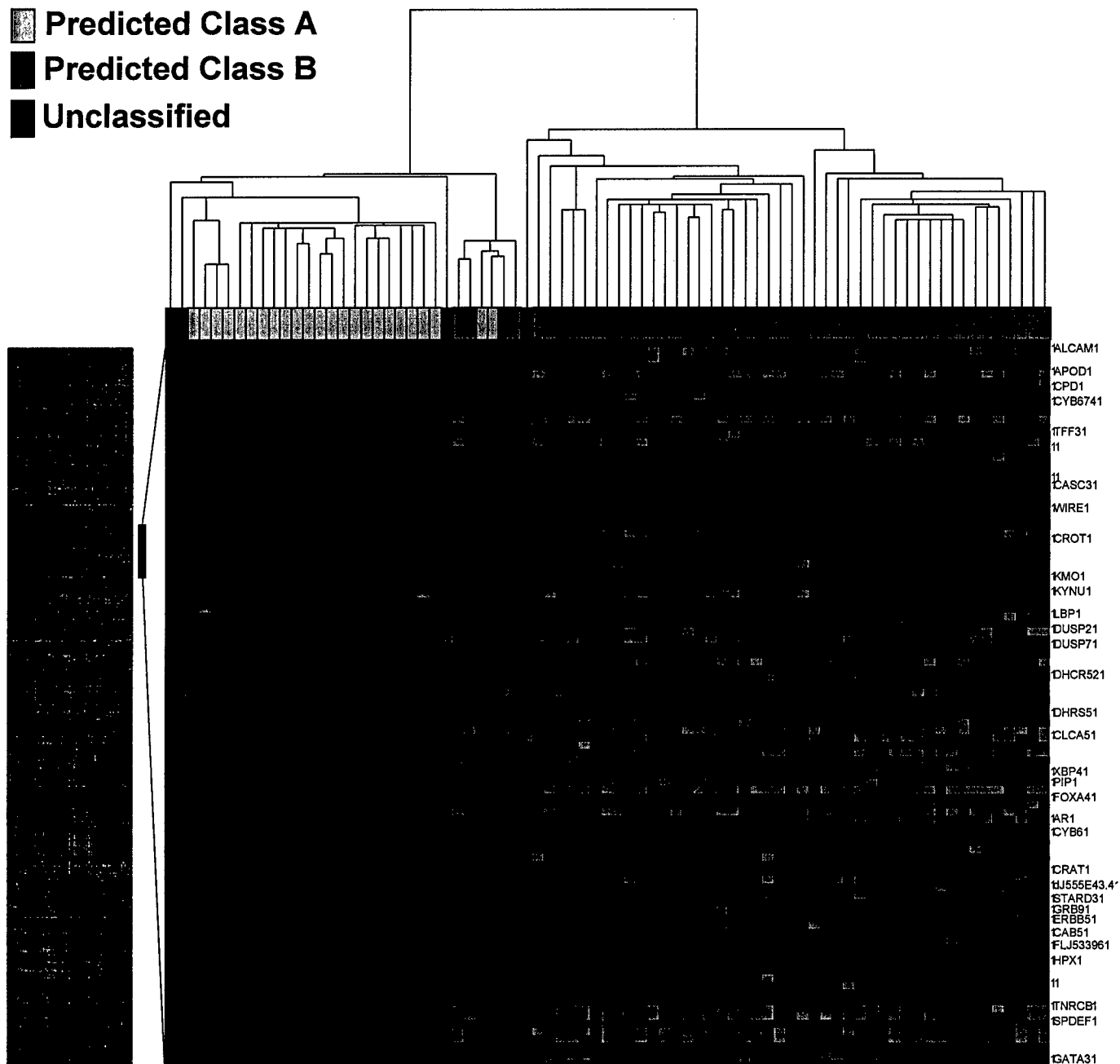


Figure 4.

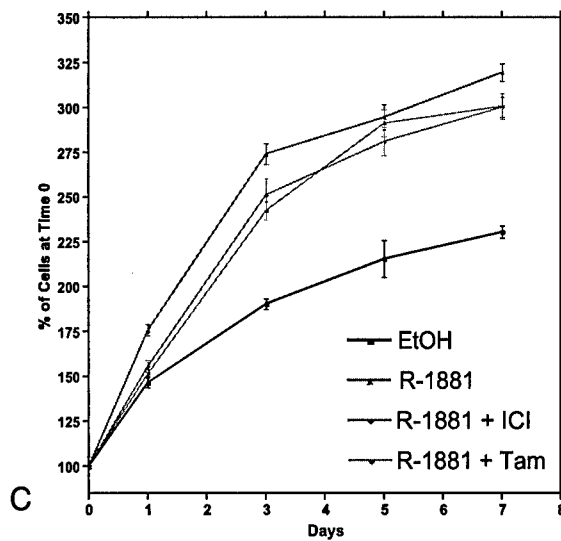
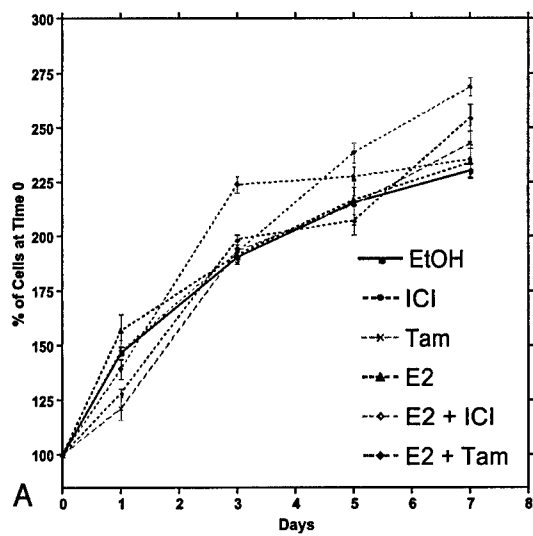
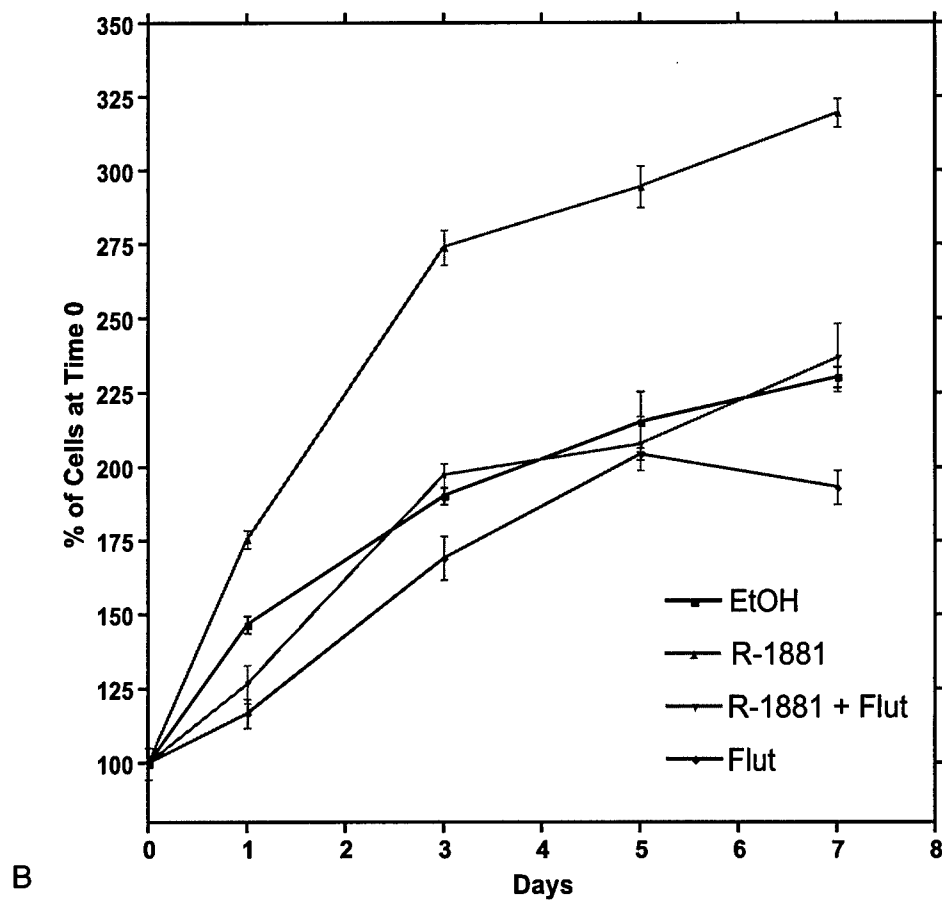
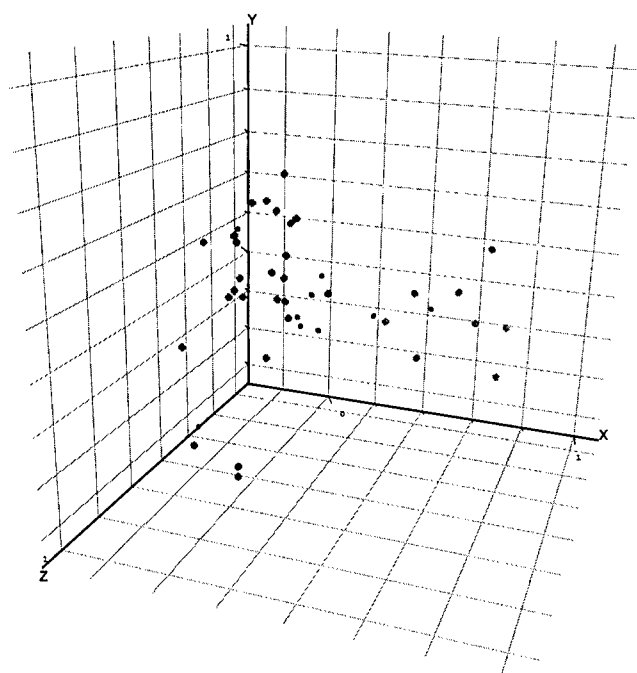


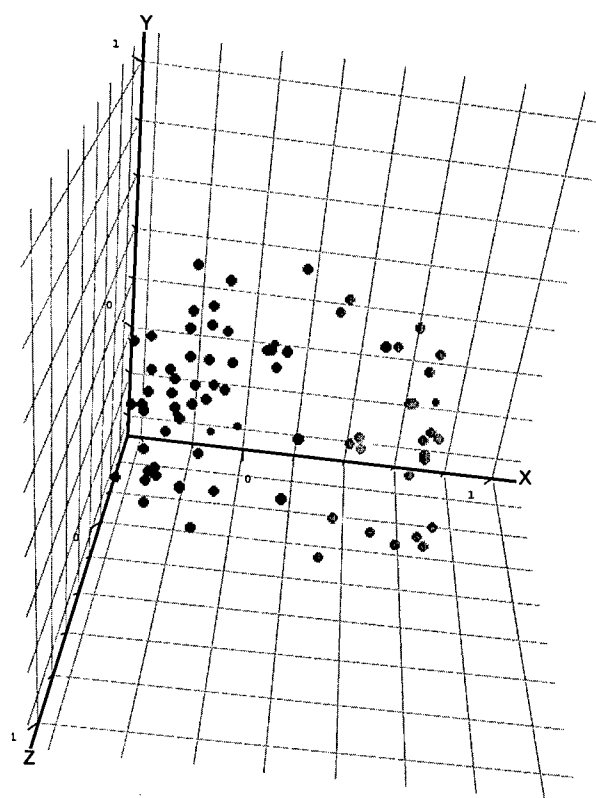
Figure 5.



X-axis: [PCA component 1 (12.55% variance)]
 Y-axis: [PCA component 2 (7.908% variance)]
 Z-axis: [PCA component 3 (6.278% variance)]

A

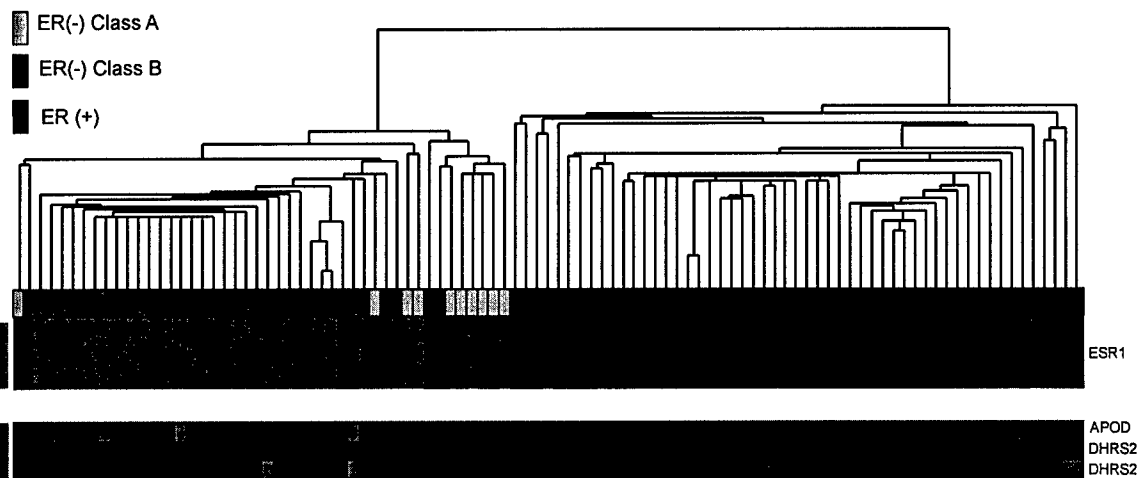
■ Class A
 ■ Class B
 ■ Unclassified



X-axis: [PCA component 1 (15.38% variance)]
 Y-axis: [PCA component 2 (6.096% variance)]
 Z-axis: [PCA component 3 (4.744% variance)]

B

Figure 6.



Supplementary Figure 1.

Nature, in press

Genes that mediate breast cancer metastasis to lung

Andy J. Minn ^{*1,2}, Gaorav P. Gupta ^{*1}, Peter M. Siegel ^{1,7}, Paula D. Bos ¹, Weiping Shu ¹,
Dilip D. Giri ^{3,8}, Agnes Viale ⁵, Adam B. Olshen ⁴, William L. Gerald ³
and Joan Massagué ^{1,6}

¹Cancer Biology and Genetics Program,
Departments of ²Radiation Oncology, ³Pathology and ⁴Epidemiology & Biostatistics,
⁵Genomics Core Laboratory, and ⁶Howard Hughes Medical Institute,
Memorial Sloan Kettering Cancer Center, New York, New York 10021, USA

* These authors made equal contributions to this work.

⁷ Present Address: McGill University Health Centre, Montreal, Quebec, Canada H34
1A4

⁸ Present Address: Department of Pathology and Laboratory Medicine, Brown
University, Providence, RI 02912

Correspondence:

Joan Massagué, Ph.D.
Cell Biology Program, Box 116
Memorial Sloan-Kettering Cancer Center
1275 York Avenue, New York, NY10021, USA

Phone: 212-639-8975 Fax: 212-717-3298 j-massague@ski.mskcc.org

By means of in vivo selection, transcriptomic analysis, functional verification and clinical validation, we identified a set of genes that marks and mediates breast cancer metastasis to lung. Some of these genes serve dual functions, providing growth advantages both in the primary tumor and in the lung microenvironment. Others contribute to aggressive growth selectively in the lung. Many encode extracellular proteins and are of previously unknown relevance to cancer metastasis.

Metastasis is frequently a final and fatal step in the progression of solid malignancies. Tumor cell intravasation, survival in circulation, extravasation into a distant organ, angiogenesis, and uninhibited growth constitute the metastatic process¹. The molecular requirements for some of these steps may be tissue-specific. Indeed, the proclivity that tumors have for specific organs, such as breast carcinomas for bone and lung, was noted over a century ago².

The identity and time of onset of the changes that endow tumor cells with these metastatic functions are largely unknown and the subject of debate. It is believed that genomic instability generates large-scale cellular heterogeneity within tumor populations, from which rare cellular variants with augmented metastatic abilities evolve through a Darwinian selection process^{2,3}. Work on experimental metastasis using tumor cell lines has demonstrated that re-injection of metastatic cell populations can enrich for the metastatic phenotype⁴⁻⁶. Recently, however, the existence of genes expressed by rare cellular variants that specifically mediate metastasis has been challenged⁷.

Transcriptomic profiling of primary human carcinomas have identified gene expression patterns that, when present in the bulk primary tumor population, predict poor patient prognosis⁸⁻¹⁰. The existence of such signatures has been interpreted to mean that genetic lesions acquired early in tumorigenesis are sufficient for the metastatic process, and that consequently no metastasis-specific genes exist⁷. However, it is unclear whether these genes that predict metastatic recurrence are also functional mediators.

The lungs and bones are frequent sites of breast cancer metastasis, and metastases to these sites differ in terms of their evolution, treatment, morbidity and mortality¹¹.

Reasoning that each organ places different demands on circulating cancer cells for the establishment of metastases, we sought to identify genes expressed in breast cancer cells that selectively mediate lung metastasis and that correlate with the propensity of primary human breast cancers to relapse to the lung.

Selection of cells metastatic to lung

The cell line MDA-MB-231 was derived from the pleural effusion of a breast cancer patient suffering from widespread metastasis years after removal of her primary tumor¹². Individual MDA-MB-231 cells grown and tested as single cell-derived progenies (SCPs) exhibit distinct metastatic ability and tissue tropism¹³ despite having similar expression levels of genes constituting a validated Rosetta-type poor prognosis signature⁹ (Supplementary Figure S1). These different metastatic behaviors, including different tropisms to bone and lung, are associated with discrete variation in overall gene expression patterns (Supplementary Figure S1; ref. ¹³). Thus, we hypothesized that organ-specific metastasis must be determined by genes that are distinct from a Rosetta-type poor prognosis signature and are differentially expressed within the MDA-MB-231 population. Indeed, previous work has demonstrated this to be the case for most of the genes linked to the activity of bone metastatic subpopulations^{4,13}.

To identify genes that mediate lung metastasis we tested parental MDA-MB-231 cells and the 1834 sub-line (an *in vivo* isolate with no enhancement in bone metastatic behavior⁴) (Figure 1a) by tail vein injection into immunodeficient mice (Figure 1b). Metastatic activity was assayed using bioluminescence imaging of luciferase-transduced cells as well as gross examination of the lungs at necropsy. The 1834 cells exhibited limited but significant lung metastatic activity compared to the parental population (Figure 1b). When 1834-derived lung lesions were expanded in culture and re-inoculated into mice, these cells (denoted as LM1 subpopulations; Figure 1a)

exhibited increased lung metastatic activity. Another round of *in vivo* selection yielded second-generation populations (denoted LM2) that were rapidly and efficiently metastatic to lung (Figure 1b). Histological analysis confirmed that LM2 lesions replaced large areas of the lung parenchyma, whereas 1834 cells exhibited intravascular growth with less extensive extravasation and parenchymal involvement (Figure 1c). Inoculation of as few as 2×10^3 LM2 cells was sufficient for the emergence of aggressive lung metastases whereas inoculation of 2×10^5 parental cells left only a residual, indolent population in the lungs (Figure 1d). Furthermore, the enhancement in lung metastatic activity was tissue-specific. When LM2 populations were inoculated into the left cardiac ventricle to facilitate bone metastasis, their metastatic activity was comparable to that of the parental and 1834 populations, and it was markedly inferior to that of a previously described, highly aggressive bone metastatic population (Figure 1b).

Elucidation of a lung metastasis signature

To identify patterns of gene expression associated with aggressive lung metastatic behavior, we performed transcriptomic microarray analysis of the highly and weakly lung metastatic cell populations. The gene list obtained from a class comparison between parental and LM2 populations was filtered to exclude genes that were expressed at low levels in a majority of samples and to ensure a 3-fold or higher change in expression level between the two groups. A total of 95 unique genes (113 probe sets) met these criteria with 48 overexpressed and 47 underexpressed in cell populations most metastatic to the lung (Figure 2a; Supplementary Table 2). This gene set was largely distinct from the bone metastasis gene-expression signature previously identified in bone metastatic isolates derived from the same parental cell line⁴. In fact, only 6 genes overlapped with concordant expression patterns between the two groups (Supplementary Table 3).

Hierarchical clustering with the 95-gene list confirmed a robust relationship between this gene expression signature and the lung-specific metastatic activity of *in vivo*-selected

cell populations (Figure 2a). In addition, this gene expression signature segregated the SCPs (which were not used in generation of the gene list) into two major groups, one transcriptomically resembling the parental cells, and the other more similar to the in vivo-selected lung metastatic populations. This latter group of SCPs was also more metastatic to lung than the former group (Figure 2b). However, unlike the LM2 populations, none of the lung metastatic SCPs concordantly expressed all of the genes in the lung metastasis signature (Figure 2a). Consistent with this observation, the lung metastatic activity of the LM2 populations was approximately one order of magnitude greater than the most aggressive SCPs (Figure 2b). We postulated that the subset of genes from the 95-gene signature that are uniformly expressed by all lung metastatic SCPs and in vivo-selected populations may confer baseline lung metastatic functions, which we define as lung metastagenicity. Genes that are expressed exclusively in the most aggressive LM2 populations may serve specialized, lung-restricted functions, which we collectively denote as lung metastatic virulence. A final list of 54 candidate lung metastagenicity and virulence genes was selected for further evaluation (Supplementary Methods and Supplementary Table 4).

Genes that mediate lung metastasis

A subset of biologically intriguing genes overexpressed in the 54 gene list was selected for functional validation. These genes include the EGF family member epiregulin (*EREG*), which is a broad-specificity ligand for the HER/ErbB family of receptors^{14,15}, the chemokine *GRO1/CXCL1*¹⁶, the matrix metalloproteinases *MMP1* (collagenase 1)¹⁷ and *MMP2* (gelatinase A)¹⁸, the cell adhesion molecule *SPARC*¹⁹, the interleukin-13 decoy receptor *IL13R α 2*²⁰ and the cell adhesion receptor *VCAM1*^{21,22} (Figure 2a). These genes encode secretory or receptor proteins, suggesting roles in the tumor cell microenvironment. In addition to these genes, we also included the transcriptional inhibitor of cell differentiation and senescence *ID1*^{23,24} and the prostaglandin-endoperoxide synthase *PTGS2/COX2*²⁵. Northern blot analysis of the various in vivo-selected cell populations revealed expression patterns for these genes that correlated

with metastatic behavior (Figure 2c). *SPARC*, *IL13R α 2*, *VCAM1* and *MMP2* belong to the subset of genes whose expression is generally restricted to aggressive lung metastatic populations and are rarely expressed (less than 10% prevalence for *VCAM1* and *IL13R α 2*, and less than 2% prevalence for *SPARC* and *MMP2*) among randomly picked SCPs (data not shown). In contrast, the expression of *ID1*, *CXCL1*, *COX2*, *EREG*, and *MMP1* is not restricted to aggressive lung metastasis populations but increases with lung metastatic ability. Analysis of protein expression for these genes confirmed that the differences in mRNA levels translated into significant alterations in protein levels (Supplementary Figure S2).

To determine if these genes play a causal role in lung metastasis, they were overexpressed via retroviral infection in the parental population either individually, in groups of three, or in groups of six (Supplementary Figure S3). Only cells overexpressing *ID1* alone were modestly more active at forming lung metastases when compared to cells infected with vector controls (Figure 3a). Consistent with the hypothesis that metastasis requires the concerted action of multiple effectors, combinations of these genes invariably led to more aggressive metastatic activity and some combinations recapitulated the aggressiveness of the 4175 LM2 population (Figure 3b). Triple combinations of lung metastasis genes in parental cells did not enhance bone metastatic activity (Supplementary Figure S4), supporting their identity as tissue-specific mediators of metastasis. The necessity of some of these genes was tested by stably decreasing their expression in 4175 (LM2) cells with short-hairpin RNAi vectors (Figure 3c). Reduction of *ID1*, *VCAM1*, or *IL13R α 2* levels decreased the lung metastatic activity of 4175 cells by more than 10-fold (Figure 3d). These effects are not due to activation of the RNAi machinery, because efficient knock down of another gene, *ROBO1*, did not inhibit lung metastasis formation (data not shown). Collectively, the results show that these nine genes are not only markers but also functional mediators of lung-specific metastasis.

The lung metastasis signature in primary tumors

A biologically meaningful and clinically relevant gene profile that mediates lung metastasis should be uniquely expressed by a subgroup of patients that relapse to the lung and it should associate with clinical outcome. To test this, a cohort of 82 breast cancer patients treated at our institution was used in a univariate Cox proportional hazards model to relate the expression level of each lung metastasis signature gene with clinical outcome. Twelve of the 54 genes are significantly associated with lung metastasis-free survival, including *MMP1*, *CXCL1*, and *PTGS2* (Supplementary Table 5). A cross-validated multivariate analysis using a linear combination of each of the 54 genes weighted by the univariate results²⁶ distinguished patients divided into a high or a low risk group for developing lung metastasis (10 year lung metastasis-free survival of 56% vs 89%, $p=0.0018$; see Supplementary Figure S5) but not bone metastasis (70% vs 79%, $p=0.31$). When a similar multivariate analysis was performed by weighting each gene by a t-statistic derived from comparing its expression between the LM2 cell lines with the parental MDA-MD-231 cells, the 54 genes again distinguished patients at high risk for developing lung metastasis (62% vs 88%, $p=0.01$; see Supplementary Figure S5) but not bone metastasis (75% vs 79%, $p=0.49$). These results suggest that a clinically relevant subgroup of patients express certain combinations of lung metastasis signature genes.

To directly determine the extent to which breast cancers express the lung metastasis signature in a manner resembling the LM2 cell lines, the 54-genes were used to hierarchically cluster the MSKCC data set. Manual inspection of branches in the dendrogram revealed a group of primary tumors that concordantly expressed many elements of this signature (Figure 4a, dashed red box). In particular, a subgroup of primary tumors expressed to varying degrees a majority of the nine genes that were functionally validated. Interestingly, many patients that developed lung metastasis were among this group. Tumors in this group predominantly expressed markers of clinically aggressive disease including negative estrogen receptor/progesterone receptor status,

a Rosetta-type poor-prognosis signature⁸, and a basal cell subtype of breast cancer²⁷. There was no association of our signature with high HER2 expression. A molecularly similar subgroup of breast cancer was identified when the clustering analysis was repeated on a previously published Rosetta microarray data set of breast cancer patients⁹ (Supplementary Figure S6), suggesting that the findings are not unique to our cohort of patients.

Although the results of the hierarchical clustering are suggestive, this approach can lead to arbitrary class assignments and is generally not ideal for class prediction²⁸.

Therefore, we took advantage of the repeated observation of our signature in two independent data sets. For training purposes the Rosetta data set was used to define a group of patients expressing the lung metastasis signature most resembling the LM2 cell lines (Supplementary Figure S7). All 48 out of the 54 lung metastasis genes that were shared between the MSKCC and Rosetta data set microarray platforms were subsequently utilized to generate a classifier to distinguish these tumors from the remaining tumors in the cohort (Supplementary Table 6). This classifier was then applied to the MSKCC cohort to identify tumors that express the lung metastasis signature in a manner resembling the LM2 cell lines. These patients had a markedly worse lung metastasis-free survival ($p < 0.001$; Figure 4b) but not bone metastasis-free survival ($p = 0.15$; Figure 4b). These results were independent of ER status and classification as a Rosetta-type poor prognosis tumor (Figure 4c). Six of the nine genes that we tested in functional validation studies (*MMP1*, *CXCL1*, *PTGS2*, *ID1*, *VCAM1*, and *EREG*) were among the 18 most univariately significant ($p < 0.05$) genes that distinguished the patients used to train the classifier (Supplementary Figure S7 cluster 3 and Table 1), and classification using only these 18 genes gave similar results (data not shown). The three remaining genes (*SPARC*, *IL13RA2*, *MMP2*) are members of the lung metastasis virulence subset and were expressed only in the most highly metastatic cell lines in our model system (Figure 2d).

Breast tumorigenicity and lung metastagenicity partially overlap

How and when metastasis genes are acquired is unknown²⁹. One explanation for the expression of a lung metastasis signature in a subgroup of primary breast cancer is that these genes may confer a growth advantage to the primary tumor while allowing growth at distant sites⁷. To test this hypothesis, MDA-MB-231 cells were orthotopically injected into the mammary fat pad of immunodeficient mice. We found that the 1834 (LM0) and 4175 (LM2) cell populations were progressively more aggressive at growing in the mammary fat pad compared with the parental cell line. This correlated with expression of lung metastagenicity genes (Figure 5a; Figure 2c) and was not due to a general enhancement of growth because the 4175, 1834, and parental populations had a comparable ability to metastasize to bone (refer to Figure 1d). Furthermore, the 4175 and 1834 populations were also more metastatic to the lungs from the orthotopic site after primary tumor resection, re-capitulating the phenotypes observed using the tail vein metastasis assay (Figure 5b). In contrast, the virulently bone metastatic population 1833⁴ was only marginally more aggressive in the mammary fat pad compared to the parental cells and did not metastasize to lung following primary tumor resection (Figures 5a and 5b).

To identify which of the genes in the lung metastasis signature may be conferring growth at the primary tumor site, we quantified mammary fat pad tumor growth of 4175 cell populations with stable knockdown of various lung metastasis genes that were previously assayed for effects on metastatic behavior (refer to Figures 3c and 3d). Whereas knockdown of IL13R α 2, SPARC, and VCAM1 decreased lung metastatic ability but not orthotopic tumor growth, knockdown of ID1 resulted in a statistically significant reduction in both (Figure 5c and Figure 3d). These data suggest that some lung metastasis genes facilitate both breast tumorigenicity and lung metastagenicity, whereas others confer growth advantages exclusively in the lung microenvironment.

Discussion

We have identified a set of genes that mediates breast cancer metastasis to lung and clinically correlates with the development of lung metastasis when expressed in primary breast cancers. Many of the genes in this signature have not been previously linked to metastasis. Together with the bone, the lung is one of the most frequent targets of breast cancer metastasis in humans. We provide evidence that these two sites impose different requirements for the establishment of metastases by circulating cancer cells. In addition to providing clinical validation, potential prognostic tools and possible targets for cancer treatment, the present findings shed new light into the biology of breast cancer metastasis.

Many of the genes in the lung metastasis signature are frequently expressed in all MDA-MB-231 subpopulations that metastasize to the lung, regardless of whether these cells were randomly picked from the parental cell line or selected *in vivo*. The majority of these genes, which we denote as promoting lung metastagenicity, encode extracellular products including growth and survival factors (e.g. the HER/ErbB receptor ligand Epiregulin), chemokines (CXCL1), cell adhesion receptors (e.g. ROBO1) and extracellular proteases (MMP1). They also include intracellular enzymes (e.g. COX2) and transcriptional regulators (e.g. ID1), as well as several intriguing downregulated genes. Their expression pattern is tightly correlated with lung metastatic activity. When tested by overexpression in poorly metastatic cells or by RNAi-mediated knockdown in highly metastatic cells, several genes in this group function as mediators of lung metastasis but not bone metastasis. Furthermore, in the cohort of human breast cancer primary tumors examined, those expressing the lung metastasis signature had a significantly worse lung metastasis-free survival but not bone metastasis-free survival. Therefore, this signature appears to include a set of clinically relevant genes that mediate a metastagenicity function^{30,31} with selectivity to the lung.

Recent data as well as our data reveal the existence of metastasis gene signatures expressed by primary tumors. It is unclear at what point these metastasis gene signatures are acquired during the process of tumorigenesis since the selection

pressure for this acquisition is unknown. One possibility is that elements of metastasis gene signatures may play a role in primary tumor growth. Consistent with this idea, the *in vivo* selected cell lines expressing the lung metastagenicity signature are more tumorigenic when implanted in the mammary glands of mice. Despite promoting growth in the mammary gland and in the lung, these genes are not general mediators of neoplastic growth. Therefore, many lung metastasis signature genes appear to enhance growth both within the breast and the lung (Figure 5d). These overlapping functions may explain how cells expressing genes involved in metastasis can be selected for in the primary tumor, providing insight into the interpretation of primary tumor microarray data.

Another subset of the lung metastasis genes is overexpressed only in rare, virulently metastatic cells selected *in vivo*. Several of these genes mediate lung metastasis in our functional assays. Many in this class encode extracellular proteins (e.g. *SPARC*, *MMP2*). With some exceptions (e.g. the receptors *IL13RA2*, *VCAM1*), this group of genes is sporadically expressed in human primary breast tumors. We propose that these genes act mainly as virulence genes^{30,31} that may allow tumors to aggressively invade, colonize, and grow in the lung without markedly contributing to primary tumor growth (Figure 5d). As such, their expression may be rare in primary tumors but strongly selected for once such cells reach the lung. Supporting this model, a recent study analyzing *MMP2* expression in matched primary breast cancers and pleural effusions found that *MMP2* levels are specifically enriched at the metastatic site³².

Breast cancer is a heterogeneous disease with diverse metastatic behavior. As a consequence, patients differ widely in prognosis and survival. Attempts to molecularly classify this disease have yielded several useful markers of poor prognosis. However, to our knowledge none of these markers have thus far been shown to act as functional mediators that account for the diversity of breast cancer metastases. In contrast, our lung metastasis signature seems to identify poor-prognosis patients who are at high risk of selectively developing lung metastasis, consistent with the functional testing done

experimentally. Further studies using additional patient cohorts and a delineation of the role of these genes in specific steps of the metastatic process, should lead to a better understanding of the biology of metastasis and its susceptibilities to treatment.

Experimental Procedures

Cell lines. The parental MDA-MB-231 cell line was obtained from the American Type Tissue Collection. Its derivative cell lines and SCPs were previously described⁴. Cells were grown in high-glucose Dulbecco's modified Eagles medium with 10% fetal bovine serum. For bioluminescent tracking, cell lines were retrovirally infected with a triple fusion protein reporter construct encoding herpes simplex virus thymidine kinase 1, green fluorescent protein (GFP) and firefly luciferase^{13,33}. GFP-positive cells were enriched by fluorescence-activated cell sorting.

Animal studies. All animal work was done in accordance with an IACUC approved protocol. Four to 6-week-old Balb/c nude mice (NCI) were used for all xenografting studies. For lung metastasis formation, 2×10^5 viable cells were washed and harvested in PBS and subsequently injected into the lateral tail vein in a volume of 0.1 mL. Endpoint assays were conducted at 15 weeks post-injection unless significant morbidity required that the mouse be sacrificed earlier. For bone metastasis, 1×10^5 cells in PBS were injected into the left ventricle of anesthetized mice (100 mg/kg Ketamine; 10 mg/kg Xylazine)⁴. Mice were imaged for luciferase activity immediately after injection to exclude any that were not successfully xenografted.

For mammary fat pad tumor assays, cells were harvested by trypsinization, washed twice in PBS and counted. Cells were then resuspended (1×10^7 cells/ml) in a 50:50 solution of PBS and Matrigel. Mice were anesthetized, a small incision was made to visualize the mammary gland and 1×10^6 cells were injected directly into the mammary fatpad. The incision was closed with wound clips and primary tumor outgrowth was monitored weekly by taking measurements of the tumor length (L) and width (W). Tumor volume was calculated as per $4/3\pi x L/2(W/2)^2$. For metastasis assays, tumors were surgically resected when they reached a tumor volume greater than 300 mm³. After resection, the mice were monitored by bioluminescent imaging for the development of metastases.

Bioluminescent imaging and analysis. Mice were anesthetized and retro-orbitally injected with 1.5 mg of D-luciferin (15 mg/mL in PBS). Imaging was completed between 2-5 minutes post-injection using a Xenogen IVIS system coupled to Living Image acquisition and analysis software (Xenogen). For BLI plots, photon flux was calculated for each mouse using a rectangular region of interest (ROI) encompassing the thorax of the mouse in a prone position. This value was scaled to a comparable background value (from a luciferin-injected mouse with no tumor cells), and then normalized to the value obtained immediately post-xenografting (day 0), so that all mice had an arbitrary starting BLI signal of 100.

RNA isolation, labeling and microarray hybridization. Methods for RNA extraction, labeling, and hybridization for DNA microarray analysis of the cell lines have been previously described⁴. For the primary breast tumor data, tissues from primary breast cancers were obtained from therapeutic procedures performed as part of routine clinical management. Samples were snap frozen in liquid nitrogen and stored at -80°C. Each sample was examined histologically using hematoxylin and eosin stained cryostat sections. Regions were manually dissected from the frozen block to provide consistent tumor cell content of greater than 70% in tissues used for analysis. All studies were conducted under MSKCC Institutional Review Board approved protocols. RNA was extracted from frozen tissues by homogenization in TRIzol reagent (GIBCO/BRL) and evaluated for integrity. Complementary DNA was synthesized from total RNA using a T7-promoter-tagged-dT primer. RNA target was synthesized by in vitro transcription and labeled with biotinylated nucleotides (Enzo Biochem, Farmingdale, NY). Labeled target was assessed by hybridization to Test3 arrays (Affymetrix, Santa Clara, CA). All gene expression analysis was carried out using HG-U133A GeneChip. Gene expression was quantitated using MAS 5.0 or GCOS (Affymetrix). All microarray data has been submitted to the Gene Expression Omnibus (GEO) under accession number GSE2603.

Statistical analysis. The Kaplan-Meier method was used to estimate survival curves and the log-rank test was used to test for differences between curves using WinSTAT (R. Fitch

Software). The site of distant metastasis for the patients in the MSKCC data set was determined from patient records. Patients with lung metastasis developed metastasis only to the lung or within months of metastasis to other sites. A detailed description of analytical methods used in the paper is provided in the Supplementary Methods section.

Descriptions of additional experimental procedures used are available in the Supplementary Methods section accompanying the paper on the Nature website.

Acknowledgements

We thank Robert Benezra, Yibin Kang, Clifford Hudis, Larry Norton, Neal Rosen and Catherine VanPoznak for invaluable insights and discussions, and Katia Manova and the staff of the Molecular Cytology Core Facility for assistance with immunohistochemistry. A.J.M is a recipient of the Leonard B. Holman Research Pathway fellowship. G.P.G. is supported by the NIH Medical Scientist Training Program grant GM07739, a fellowship from the Katherine Beineke Foundation, and the Department of Defense Breast Cancer Research Program pre-doctoral traineeship award W81XWH-04-1-0334. J.M. is an Investigator of the Howard Hughes Medical Institute. This research is supported by the W.M. Keck Foundation and NIH grant P01-CA94060 to JM, and U.S. Army Medical Research grant DAMD17-02-0484 to WG.

Table 1. Partial List of Lung Metastasis Signature Genes Used to Classify Primary Breast Cancers Expressing the Lung Metastasis Signature.

p-value	UG cluster	Gene symbol	Description
<0.000001	Hs.118400	FSCN1	Fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)
<0.000001	Hs.83169	MMP1	Matrix metalloproteinase 1 (interstitial collagenase)
<0.000001	Hs.9613	ANGPTL4	Angiopoietin-like 4
0.000006	Hs.74120	C10orf116	Chromosome 10 open reading frame 116
0.00002	Hs.789	CXCL1	Chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
0.000355	Hs.196384	PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
0.000444	Hs.185568	KRTHB1	Keratin, hair, basic, 1
0.000506	Hs.109225	VCAM1	Vascular cell adhesion molecule 1
0.000627	Hs.17466	RARRES3	Retinoic acid receptor responder (tazarotene induced) 3
0.001263	Hs.368256	LTBP1	Latent transforming growth factor beta binding protein 1
0.004365	Hs.444471	KYNU	Kynureninase (L-kynurenine hydrolase)
0.005179	Hs.421986	CXCR4	Chemokine (C-X-C motif) receptor 4
0.006426	Hs.77667	LY6E	Lymphocyte antigen 6 complex, locus E
0.007153	Hs.410900	ID1	Inhibitor of DNA binding 1, dominant negative helix-loop-helix protein
0.010871	Hs.255149	MAN1A1	Mannosidase, alpha, class 1A, member 1
0.032361	Hs.388589	NEDD9	Neural precursor cell expressed, developmentally down-regulated 9
0.03713	Hs.115263	EREG	Epiregulin
0.046859	Hs.98998	TNC	Tenascin C (hexabrachion)

There are 48 unique genes shared between MSKCC and Rosetta microarray platforms. Patients from the Rosetta training set were used to define a class label for patients that either express or do not express the lung metastasis signature. Shown is the p-value of a t-test comparing the difference in gene expression between these two classes (Supplementary Figure S7, cluster 3). Only 18 genes with a p-value < 0.05 are shown.

Figure 1. Selection of breast cancer cells metastatic to lung. **a**, Flow chart of the in-vivo selection of organ-specific metastatic subpopulations indicating the organs from which these subpopulations were isolated. Each subsequent lung metastatic generation is designated LM0, LM1, and LM2. The LM2 cells were further analyzed for metastasis by either tailvein (TV) or intracardiac (IC) xenografting. Metastatic propensities for all cell lines used in this study are listed in Supplementary Table 1. **b**, Representative lungs harvested at necropsy and bioluminescence imaging of the indicated cell lines are shown after tailvein or intracardiac injection. **c**, Hematoxylin staining of frozen sections of lungs from mice injected with moderately metastatic 1834 cells show a mix of invading lesions (*asterisk*) and emboli within the vascular space (*arrowheads*). Vascular walls are stained with the endothelial cell marker CD31. **d**, The indicated numbers of parental cells and 4175 (LM2) cells were tested for lung metastatic activity. Plots show a quantitation of the luminescence signal as a function of time. Data are the average \pm SEM for each cohort. (*) $p < 0.05$ using a one-sided rank test, compared to mice injected with an equivalent number of Parental cells..

Figure 2. A gene-expression signature associated with lung metastasis. **a**, Comparison of gene expression profiles of LM2 populations with parental cells identifies 113 probe sets that correlate with lung metastatic activity. This signature clusters in-vivo selected populations and single cell-derived progenies (SCPs) into groups that resemble the LM2 cell lines (*red bar*), the parental MDA-MB-231 cell line (*green bar*), or an intermediate group (*blue bar*). **b**, LM2 populations 4175 and 4142 were assayed for lung metastatic activity as measured by BLI and compared to parental populations and various SCPs¹³. Plots show a quantitation of the luminescence signal as a function of time. Data are the average \pm SEM for each cohort. Color-coding is as in panel **a**. **c**, Northern blot analysis of parental, LM0, LM1, and LM2 cell lines using a set of nine lung metastasis genes selected for functional validation, as well as four intriguing genes underexpressed in the lung metastatic populations.

Figure 3. Genes in the expression signature mediate lung metastasis. **a,b**,

Retrovirus-mediated expression of selected genes from the lung metastasis signature in weakly metastatic parental MDA-MB-231 cells. Genes were tested individually (*a*) or in groups of 3 or 6 genes (*b*). *c*, Stable short hairpin RNAi constructs were retrovirally introduced into 4175 lung metastatic cells, and their effectiveness at knocking down the expression of their intended target was validated at the protein level (ID1, VCAM1, SPARC) or mRNA level (*IL13RA2*). *d*, 4175 knockdown cell lines were xenografted via the tail vein to assess lung metastatic activity. One shRNA vector against ID1 that was ineffective at decreasing expression of this gene serves as a negative control. Data are the average \pm SEM for each cohort. (*) $p < 0.05$ using a one-sided rank test.

Figure 4. The lung metastasis signature in human primary breast tumors. a, Hierarchical clustering of primary breast carcinomas from a cohort of 82 breast cancer patients was performed using the 54 lung metastasis signature genes. A dendrogram of the tumors is shown at the top, with tumors from patients that developed lung metastasis (black circles) or non-pulmonary sites (yellow circles) denoted. A sub-cluster with a reproducibility index of 0.71 (*dashed red box*) groups tumors that tended to express the lung metastasis signature in a manner resembling the LM2 cell lines. The genes were also clustered and gene names are on the right. Functionally validated genes are in red. The Rosetta poor prognosis signature is displayed with the genes underexpressed (*green bar*) and overexpressed (*red bar*) in poor prognosis tumors indicated on the left. The expression of HER2, progesterone receptor (*PR*), estrogen receptor (*ER*), and basal and luminal keratins are shown. Expression of the lung metastasis signature was confirmed in the independent Rosetta breast cancer cohort (Supplementary Figure S6). **b,** Lung metastasis-free survival and bone metastasis-free survival for MSKCC patients that either express (red line) or do not express (blue line) the lung metastasis signature based on a classifier trained using the Rosetta cohort (Supplementary Figure S7 and Supplementary Methods). The p-value for each survival curve is shown. **c,** Lung metastasis-free survival restricted to patients with ER-negative tumors or Rosetta-type poor prognosis tumors.

Figure 5. Breast tumorigenicity and lung metastagenicity partially overlap. a,

Representative MDA-MB-231 cell populations were injected into the mammary fat pad of immunodeficient mice and monitored for tumor growth. Each curve designates mean tumor volumes in cubic millimeters \pm SEM. The number of mice in each cohort (n) is indicated. **b,** As depicted in the schematic, mice were inoculated with the indicated MDA-MB-231 cells into the mammary fat pad and tumors were removed after reaching 300 mm³. Lung metastasis was monitored with BLI and normalized photon flux was measured two weeks after removal of the primary tumor. (*) A mouse in the 4175 cohort with an unusually high signal of 36400 was excluded. **c,** Growth in mammary fat pad of highly lung metastatic 4175 (LM2) cells after stable shRNA knockdown of the indicated genes. shControl refers to a cell line transduced with a short hairpin construct that did not result in effective knockdown of its target gene. (**) $p < 0.01$ by a one-sided rank test. **d,** A model of two classes of genes contained within the lung metastasis signature. The first class (Subset A) confers both breast tumorigenicity and basal lung metastagenicity. Examples may include *ID1*, *CXCL1*, *PTGS2*, and *MMP1*. The second class (Subset B) confers functions specific to the lung microenvironment, facilitating lung metastatic virulence. Examples may include *SPARC* and *MMP2*.

References

1. Chambers, A. F., Groom, A. C. & MacDonald, I. C. Dissemination and growth of cancer cells in metastatic sites. *Nat Rev Cancer* 2, 563-72 (2002).
2. Fidler, I. J. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat Rev Cancer* 3, 453-8 (2003).
3. Yokota, J. Tumor progression and metastasis. *Carcinogenesis* 21, 497-503 (2000).
4. Kang, Y. et al. A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* 3, 537-49 (2003).
5. Clark, E. A., Golub, T. R., Lander, E. S. & Hynes, R. O. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* 406, 532-5 (2000).
6. Yang, J. et al. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* 117, 927-39 (2004).
7. Bernards, R. & Weinberg, R. A. A progression puzzle. *Nature* 418, 823 (2002).
8. van de Vijver, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347, 1999-2009 (2002).
9. van 't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-6 (2002).
10. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 33, 49-54 (2003).
11. Solomayer, E. F., Diel, I. J., Meyberg, G. C., Gollan, C. & Bastert, G. Metastatic breast cancer: clinical course, prognosis and therapy related to the first site of metastasis. *Breast Cancer Res Treat* 59, 271-8 (2000).
12. Cailleau, R., Olive, M. & Cruciger, Q. V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* 14, 911-5 (1978).
13. Minn, A. J. et al. Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J Clin Invest* 115, 44-55 (2005).
14. Shelly, M. et al. Epiregulin is a potent pan-ErbB ligand that preferentially activates heterodimeric receptor complexes. *J Biol Chem* 273, 10496-505 (1998).
15. Yarden, Y. & Sliwkowski, M. X. Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol* 2, 127-37 (2001).
16. Balkwill, F. Cancer and the chemokine network. *Nat Rev Cancer* 4, 540-50 (2004).
17. Egeblad, M. & Werb, Z. New functions for the matrix metalloproteinases in cancer progression. *Nat Rev Cancer* 2, 161-74 (2002).
18. Duffy, M. J., Maguire, T. M., Hill, A., McDermott, E. & O'Higgins, N. Metalloproteinases: role in breast carcinogenesis, invasion and metastasis. *Breast Cancer Res* 2, 252-7 (2000).
19. Framson, P. E. & Sage, E. H. SPARC and tumor growth: where the seed meets the soil? *J Cell Biochem* 92, 679-90 (2004).
20. Wood, N. et al. Enhanced interleukin (IL)-13 responses in mice lacking IL-13 receptor alpha 2. *J Exp Med* 197, 703-9 (2003).
21. Amatschek, S. et al. Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes. *Cancer Res* 64, 844-56 (2004).

22. O'Hanlon, D. M. et al. Soluble adhesion molecules (E-selectin, ICAM-1 and VCAM-1) in breast carcinoma. *Eur J Cancer* 38, 2252-7 (2002).
23. Desprez, P. Y., Sumida, T. & Coppe, J. P. Helix-loop-helix proteins in mammary gland development and breast cancer. *J Mammary Gland Biol Neoplasia* 8, 225-39 (2003).
24. Ruzinova, M. B. & Benezra, R. Id proteins in development, cell cycle and cancer. *Trends Cell Biol* 13, 410-8 (2003).
25. Arun, B. & Goss, P. The role of COX-2 inhibition in breast cancer treatment and prevention. *Semin Oncol* 31, 22-9 (2004).
26. Beer, D. G. et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 8, 816-24 (2002).
27. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* 406, 747-52 (2000).
28. Simon, R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br J Cancer* 89, 1599-604 (2003).
29. Hynes, R. O. Metastatic potential: generic predisposition of the primary tumor or rare, metastatic variants-or both? *Cell* 113, 821-3 (2003).
30. Heimann, R. & Hellman, S. Clinical progression of breast cancer malignant behavior: what to expect and when to expect it. *J Clin Oncol* 18, 591-9 (2000).
31. Schairer, C., Mink, P. J., Carroll, L. & Devesa, S. S. Probabilities of death from breast cancer and other causes among female breast cancer patients. *J Natl Cancer Inst* 96, 1311-21 (2004).
32. Davidson, B. et al. Altered expression of metastasis-associated and regulatory molecules in effusions from breast cancer patients: a novel model for tumor progression. *Clin Cancer Res* 10, 7335-46 (2004).
33. Ponomarev, V. et al. A novel triple-modality reporter gene for whole-body fluorescent, bioluminescent, and nuclear noninvasive imaging. *Eur J Nucl Med Mol Imaging* 31, 740-51 (2004).
34. Brummelkamp, T. R., Bernards, R. & Agami, R. Stable suppression of tumorigenicity by virus-mediated RNA interference. *Cancer Cell* 2, 243-7 (2002).
35. McShane, L. M. et al. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 18, 1462-1469 (2002).
36. Wilson, C. A. & Dering, J. Recent translational research: microarray expression profiling of breast cancer--beyond classification and prognostic markers? *Breast Cancer Res* 6, 192-200 (2004).
37. Stein, D. et al. The SH2 domain protein GRB-7 is co-amplified, overexpressed and in a tight complex with HER2 in breast cancer. *Embo J* 13, 1331-40 (1994).
38. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 100, 8418-23 (2003).
39. Saeed, A. I. et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374-378 (2003).
40. Yeung, K. Y., Haynor, D. R. & Ruzzo, W. L. Validating clustering for gene expression data. *Bioinformatics* 17, 309-18 (2001).
41. Cheadle, C., Vawter, M. P., Freed, W. J. & Becker, K. G. Analysis of microarray data using Z score transformation. *J Mol Diagn* 5, 73-81 (2003).

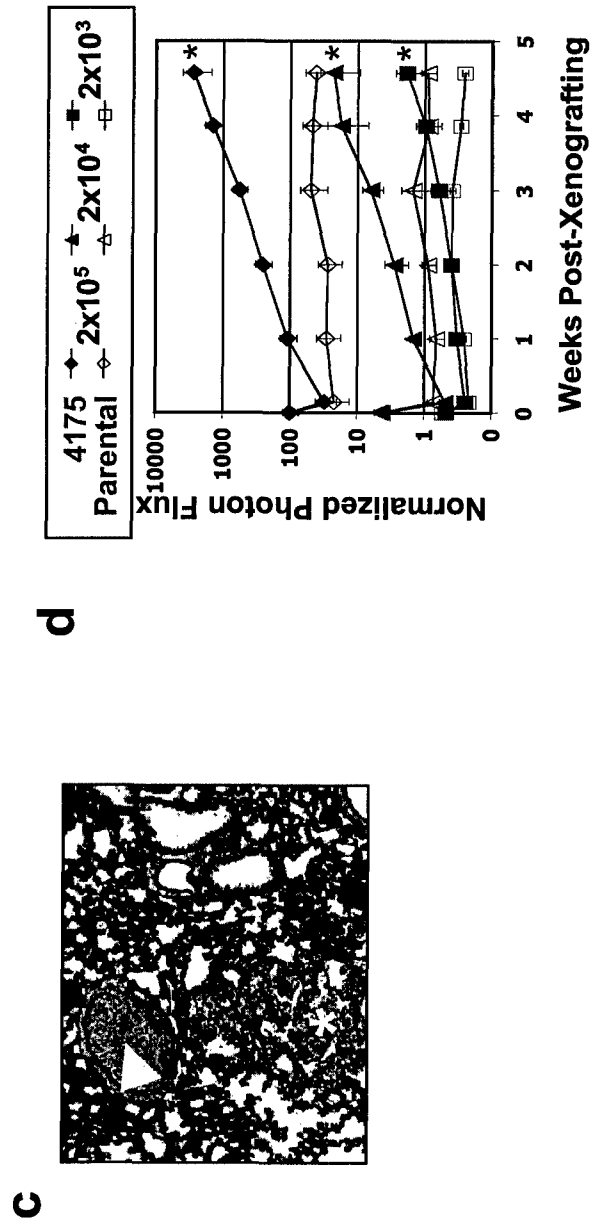
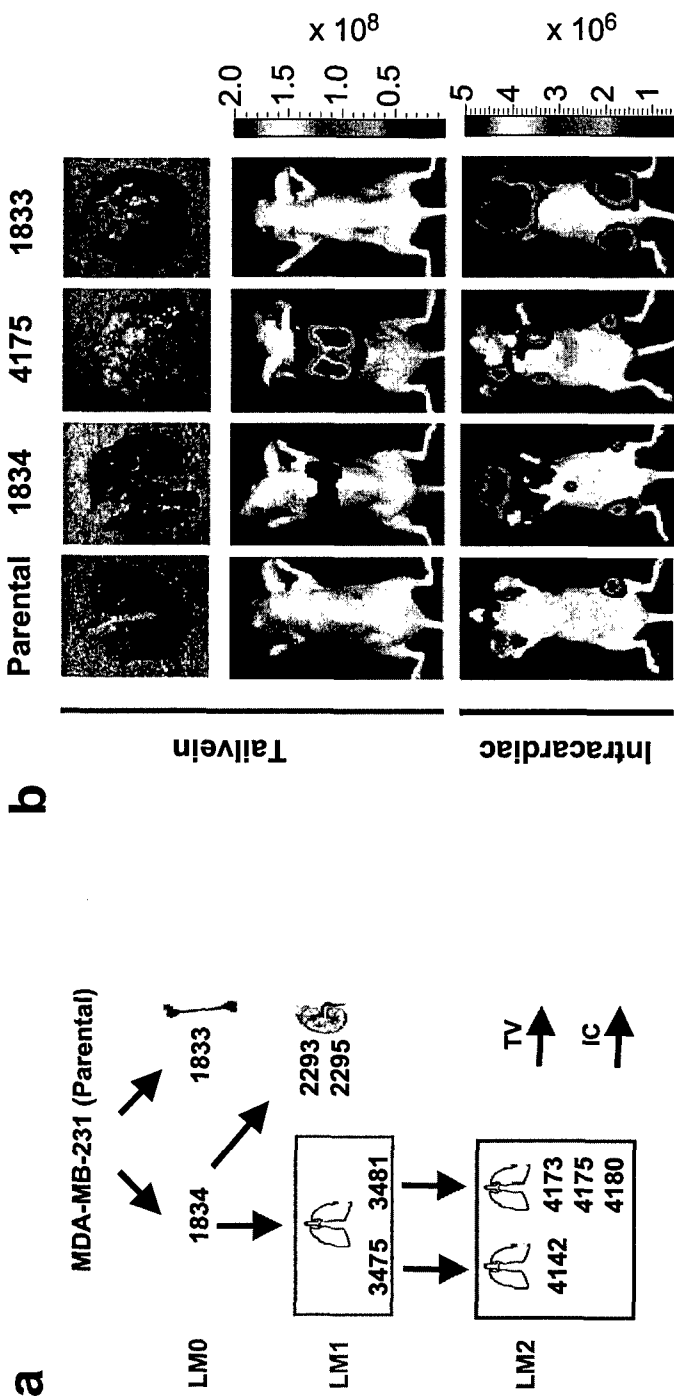


Figure 1

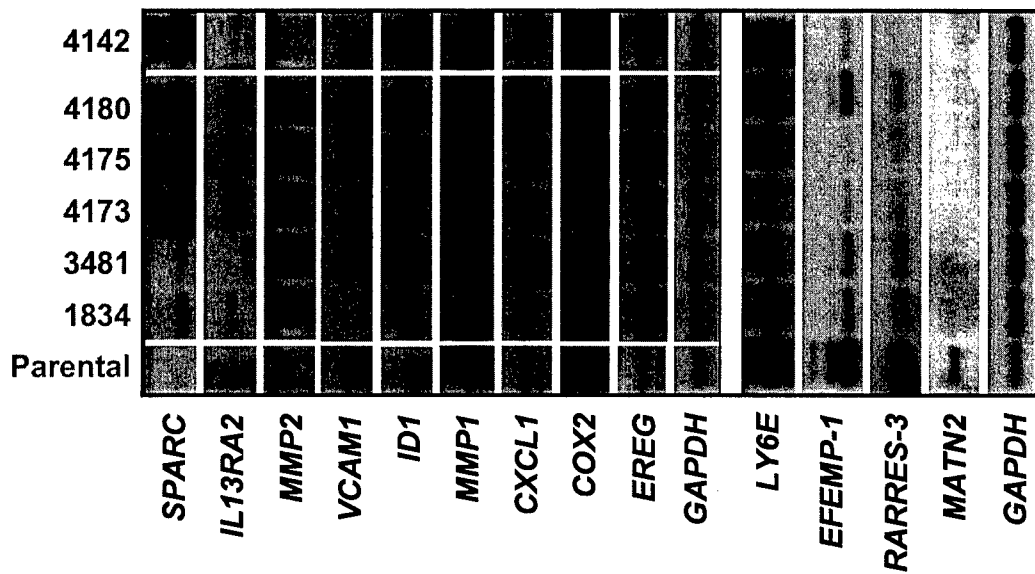
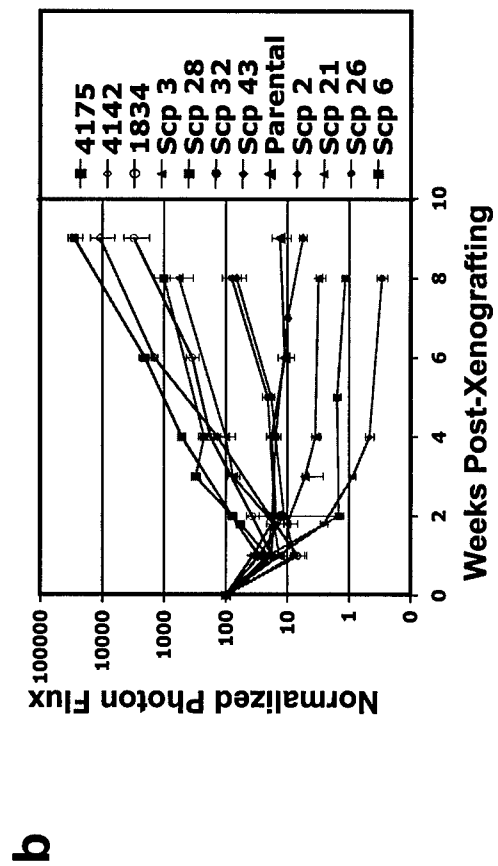
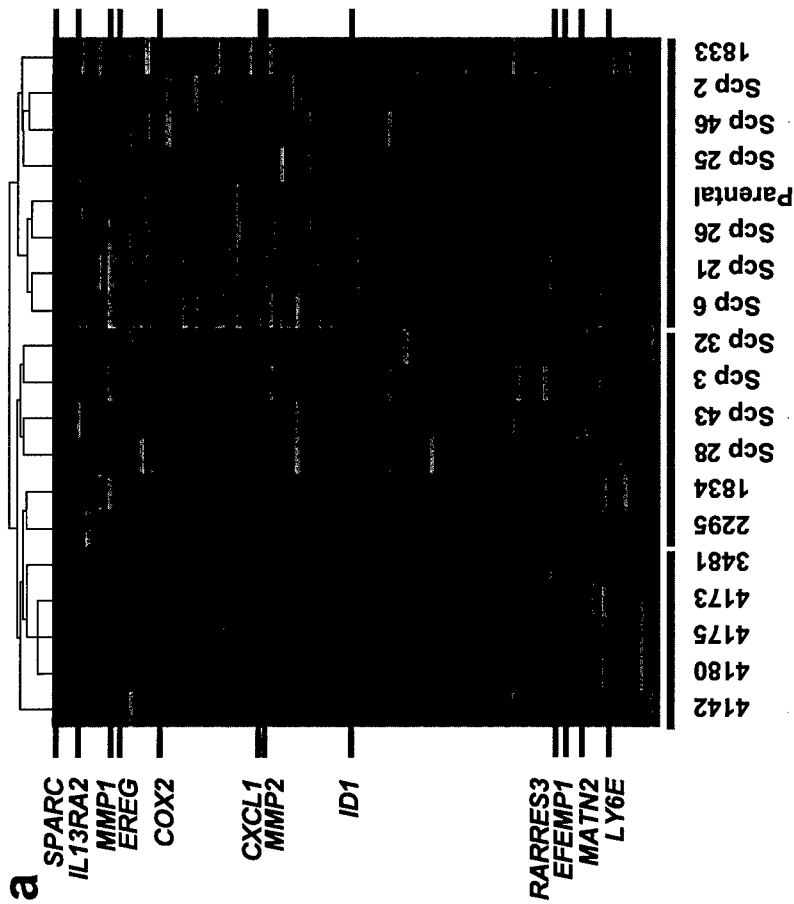


Figure 2

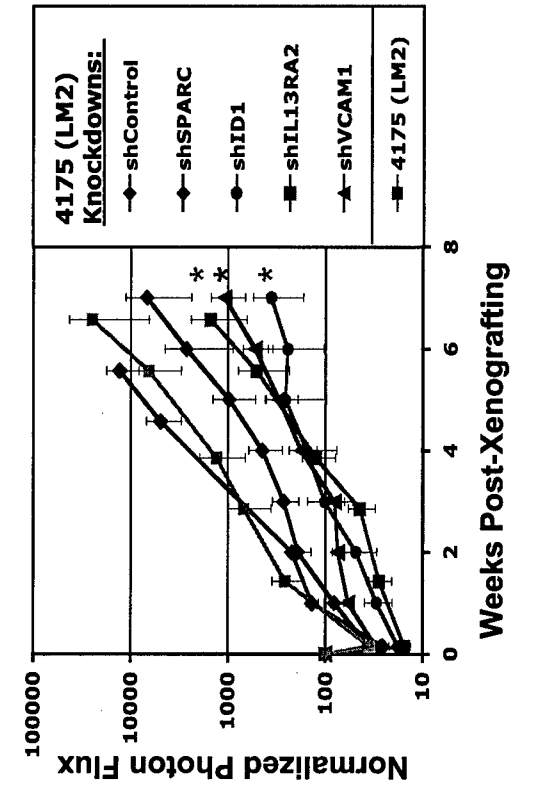
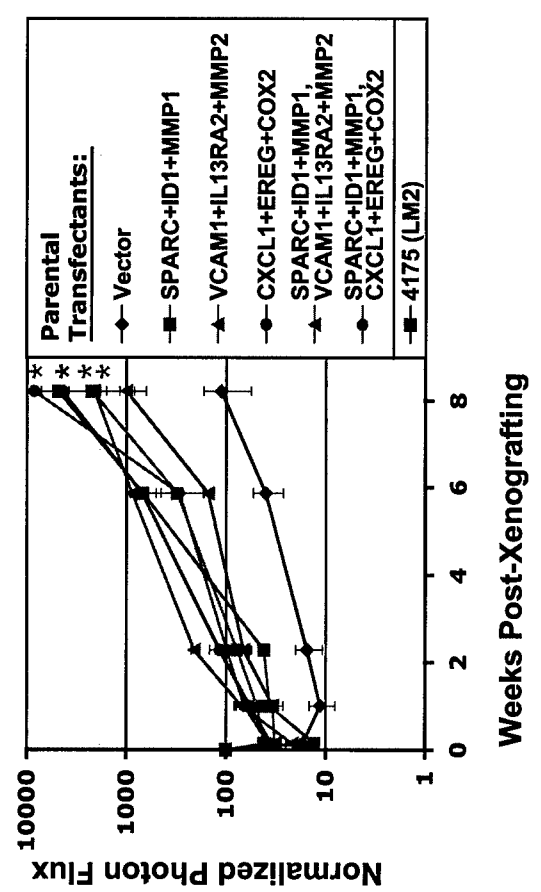
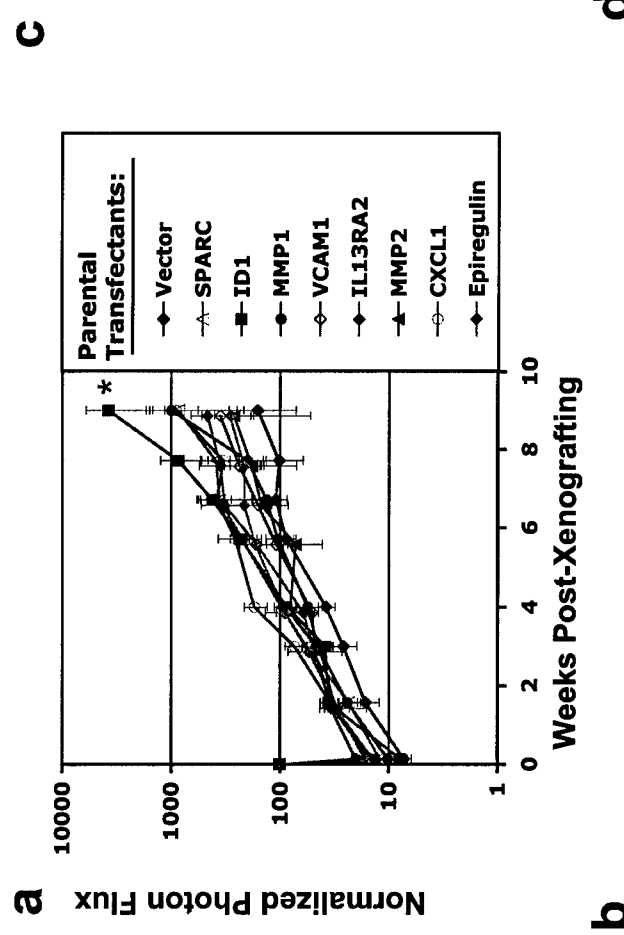
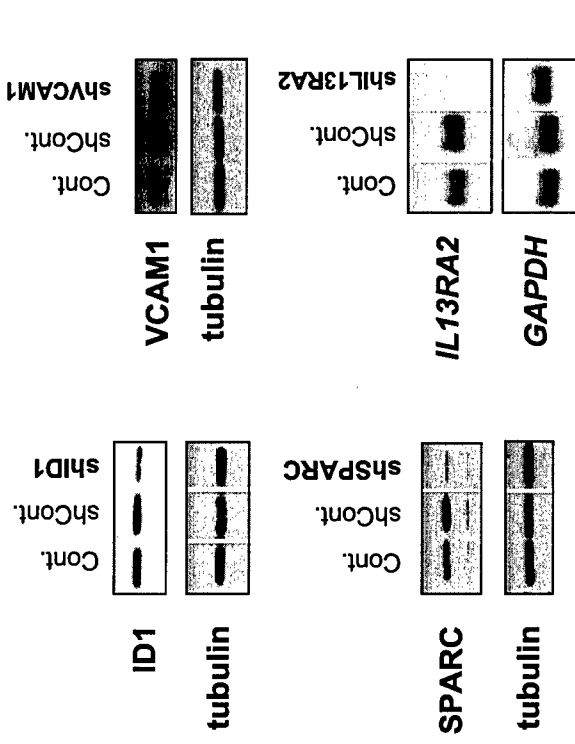
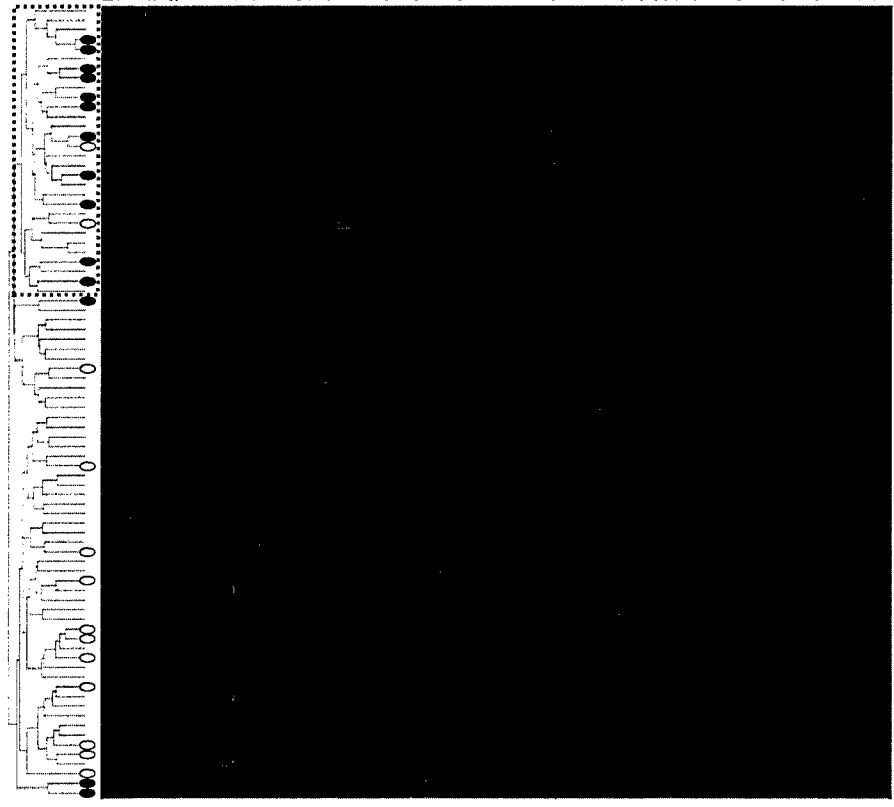


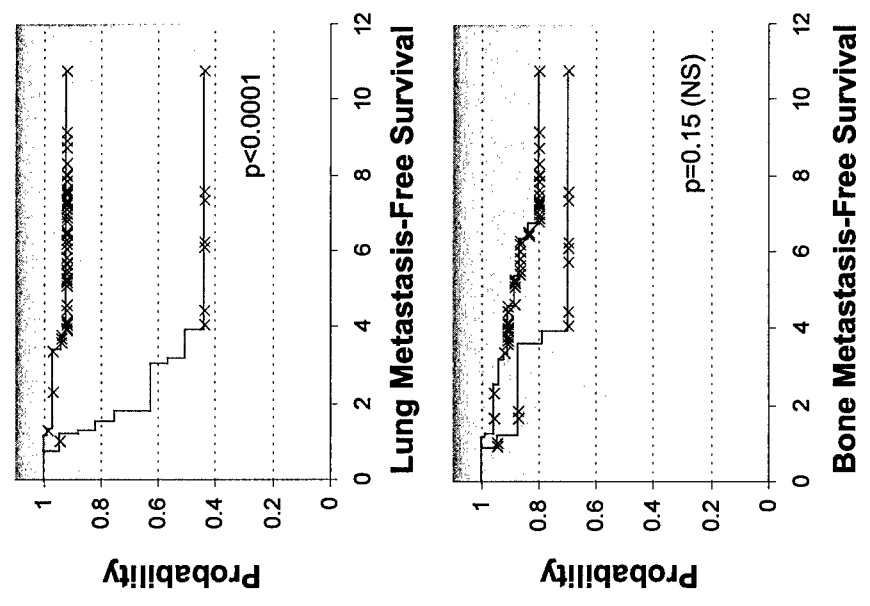
Figure 3

a



Lung Metastasis Signature

b



c

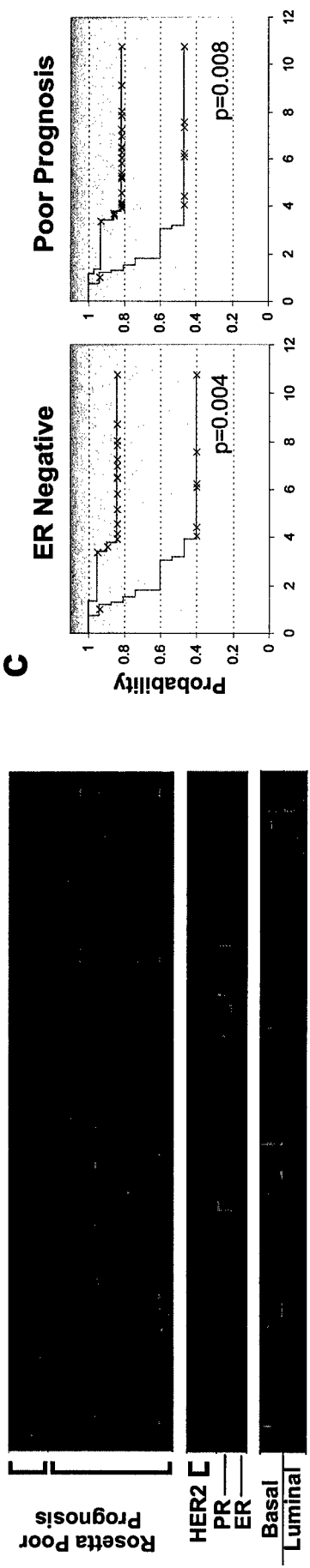
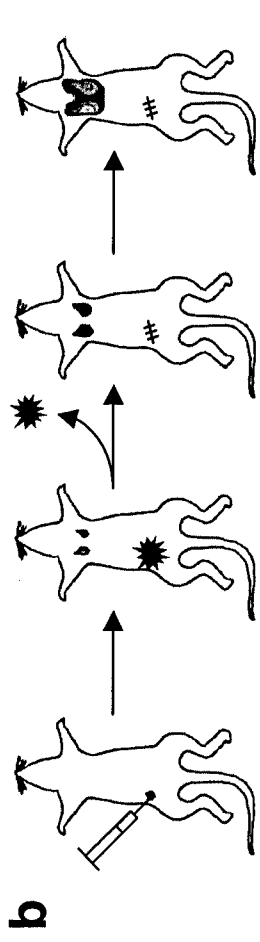
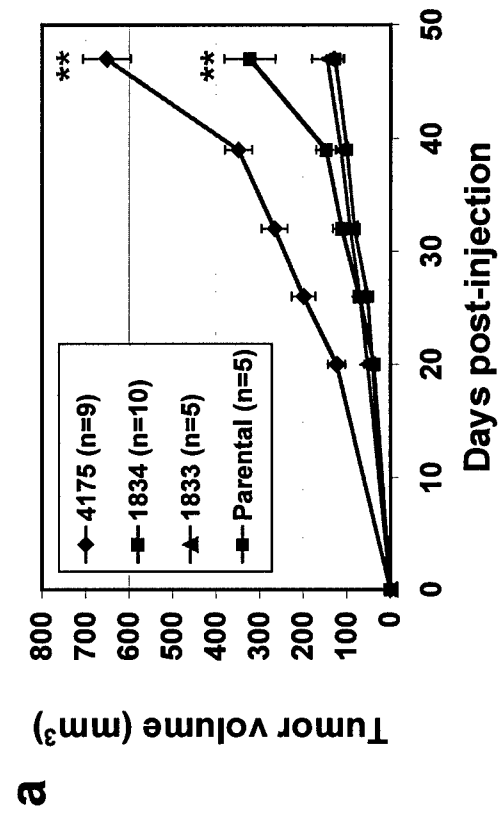


Figure 4



Cell Line	Tropism from circulation		Lung Met Frequency	Mean Intensity of Lung Met Signal
	Lung	Bone		
Parental	+	+	0/5	N/A
1833	+	+++	0/8	N/A
1834	++	+	2/7	6.5 +/- 0.03
4175	+++	+	7/13	98.3 +/- 29.3 *

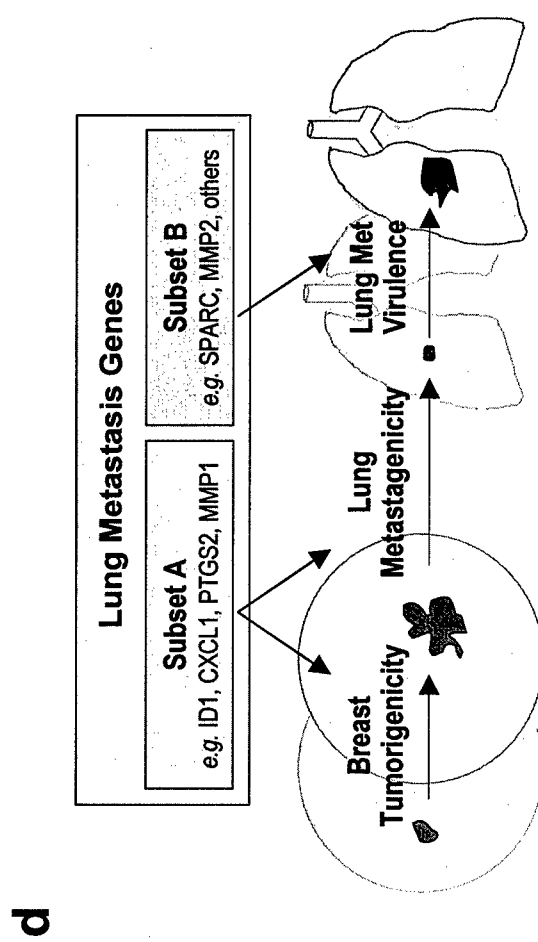
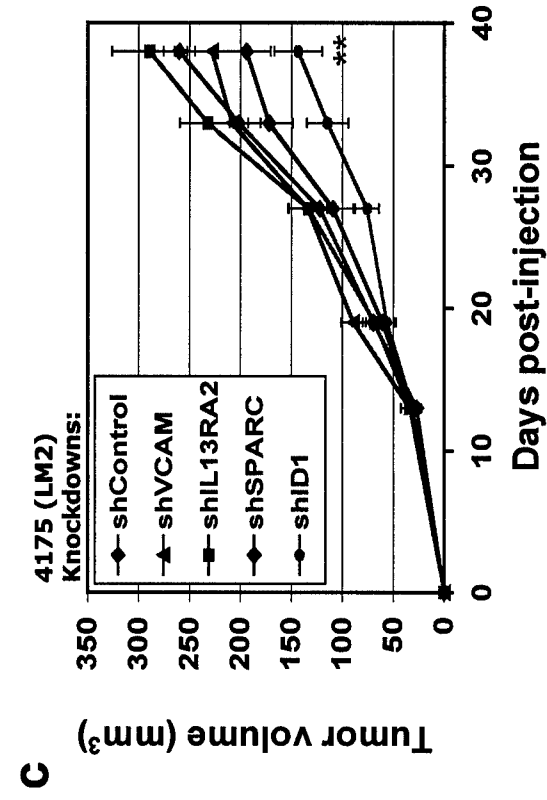


Figure 5

Supplementary Methods 1: additional experimental procedures used

Lung histology. Lungs were harvested at necropsy. For hematoxylin and eosin staining, lungs were fixed in 10% neutral buffered formalin overnight, washed with PBS and dehydrated in 70% ethanol before paraffin embedding (Histoserv). For CD31 staining, lungs were fixed in 4% paraformaldehyde overnight and treated with 30% sucrose for 12-24 h before cryosectioning. Staining was performed using anti-CD31 antibody (sc-1506, Santa Cruz Biotechnology).

Analysis of mRNA and protein expression. Total RNA from subconfluent MDA-MB-231 cells were harvested using the RNeasy kit (Qiagen). Samples were electrophoresed in MOPS buffer and transferred to a Hybond N+ membrane (Amersham). Radioactive probes for Northern blotting were derived from fragments of the relevant cDNA, and hybridization was done at 68°C for 3 h. For immunoblotting, cells were washed with PBS and lysed in RIPA buffer (50 mM Tris-HCl pH 7.4, 1% NP-40, 0.25% Na-deoxycholate, 150 mM NaCl, 1 mM EDTA) supplemented with 50 mM NaF, 20 mM β -glycerophosphate, and complete protease inhibitor cocktail (Roche). Proteins were separated by SDS-PAGE, and transferred to PVDF membranes that were immunoblotted with antibodies against ID1 or VCAM1 (Santa Cruz Biotechnology), SPARC (R&D Systems), and α -tubulin (Sigma). Secreted MMP-1, MMP-2 and CXCL1 were analyzed in conditioned media using commercially available ELISA kits (R&D Systems). Cells were plated in triplicate at 90% confluency in 6 well plates, and conditioned media was collected 48 h later. Media was cleared of cells by centrifuging at 2000 rpm for 5 min, and subsequently assayed for protein concentration according to the protocols for the relevant ELISA kits.

Cell-surface IL13R α 2 and VCAM1 were analyzed by flow cytometry in cells harvested with trypsin-EDTA and washed twice with cold PBS. CyChrome-conjugated anti-human VCAM1 (BD Pharmingen), phycoerythrin-conjugated anti-human IL13R α 2 (Cell Sciences), or control IgG were incubated in FACS buffer (0.1% sodium azide and 1% bovine serum albumin in PBS) at concentrations recommended by the supplier, for 1 h at 4 °C in the dark. Cells were washed twice and re-suspended in cold FACS buffer. Flow cytometry data was collected on a FACScalibur (BD) instrument and analyzed using FlowJo software.

Overexpression and knockdown constructs. For overexpression studies, human cDNAs of interest were cloned into pBabe-puro and/or pBabe-hygro retroviral expression vectors. For single transductions, 20 μ g of DNA were transfected into the amphotropic GPG29 packaging cell line using Lipofectamine 2000 (Invitrogen) at a ratio of 1:3 (μ g DNA: μ l Lipofectamine 2000). Virus-containing supernatants were harvested daily between 48 and 96 h post-transfection. Media was centrifuged at 2000 rpm for 5 minutes and subsequently cleared of remnant cells using a 0.45 μ m syringe filter (VWR). Filtered viral media was added to 70% confluent MDA-MB-231 cells in the presence of 8 μ g/ml polybrene (Sigma), and incubated overnight. 72 h post-infection, cell populations were treated with either puromycin (Sigma) or hygromycin (Calbiochem). Expression of the relevant transgenes was validated by Northern blot or protein expression analysis.

For combination overexpression experiments, groups of three genes expressing the same drug resistance marker were co-transfected into GPG29 packaging cells as

described, but using 15 micrograms of each plasmid. Viral harvesting and infection was identical to that described above. Sextet transductions were generated as two sequential triple infections. Cells were selected for the first drug resistance marker before being infected and selected for the second resistance marker. The *SPARC*, *ID1*, and *MMP1* triplet encoded a puromycin-resistance marker, whereas the *VCAM1*, *IL13RA2*, and *MMP2* as well as the *CXCL1*, *EREG*, and *COX2* triplets delivered hygromycin-resistant markers into the recipient cells.

For knockdown experiments, short hairpin RNAi constructs were cloned into the pRetroSuper plasmid according to previously published protocols³⁴. Retroviral infection into 4175 cells was achieved as described above for the overexpression constructs. Multiple hairpin constructs were screened for effective knockdown of the gene product of interest. 19 nucleotide target sequences that resulted in productive knockdown included: 5'-ggatctgtgatctaaatc-3' (*SPARC*), 5'-gaggaattacgtgctctgt-3' (*ID1*), 5'-ggtgaagacctatcgaaga-3' (*IL13RA2*). For knockdown of *VCAM1*, 4175 cells were sequentially infected and puromycin-selected with two different pRetroSuper targeting constructs, encoding 5'-ggcagagtacgcaaact-3' and 5'-gtccctggaaaccaagagt-3', respectively. Negative control cell lines were generated by infecting with a pRetroSuper construct targeting 5'-cggctgttactcacgcctc-3', a sequence in the *ID1* cDNA that did not yield any appreciable knockdown of the protein product by Western blotting.

Supplementary Methods 2: description of analytical methods used

Microarray data analysis of MDA-MB-231 cell lines

Analysis of transcriptomic profile heterogeneity within MDA-MB-231 human breast cancer cell line was performed using multidimensional scaling (MDS) of single cell-derived progenies (SCPs) by importing the Affymetrix data into BRBArray Tools 3.2 (Developed by R. Simon and A.P. Lam, <http://linus.nci.nih.gov/BRB-ArrayTools.html>). A list of 1267 genes that was differentially expressed across the SCPs¹³ was then used in MDS to separate the SCPs based on Pearson correlation as the similarity measure.

To identify genes associated with lung metastasis among the in vivo selected MDA-MB-231 cells, class labels were assigned based on lung metastatic behavior. A class comparison using a t-test (GeneSpring 6.1) was done between the gene expression data for second generation in vivo selected lung metastatic populations (LM2) 4173, 4175, and 4180 compared with two different passages of parental MDA-MB-231 cells (ATCC) to generate an initial list of genes that are differentially expressed between the two classes with a p-value less than 0.05. The data was further filtered to eliminate absent genes or genes expressed at low levels. This was done by removing genes with an absent flag in all the samples and genes with a raw expression score of less than 200. An additional filter was applied to ensure at least a three fold change in expression level between LM2 and ATCC, resulting in a final list consisting of 113 gene probe sets (corresponding to 95 unique genes) associated with lung metastasis. Using

these genes, hierarchical clustering was performed on a cohort of in vivo selected cell lines in addition to the SCPs, which were not directly used in the initial class comparison.

The partial expression of the 113 gene profile by the moderately lung metastatic SCPs suggests that this 113 gene list contains genes associated with baseline lung metastatic ability (lung metastagenicity), and genes that enhance this baseline behavior (lung metastatic virulence). We reasoned that lung metastagenicity genes should be differentially expressed by both the LM2 populations and the lung metastatic SCPs. Thus, a list of 59 candidate lung metastagenicity genes (50 unique) was generated by taking the intersection of the 113 genes with the 1267 genes differentially expressed across the SCPs. We reasoned that the remaining genes either represent virulence genes restricted to the most aggressive lung metastatic populations or represent false discoveries. To help distinguish between these possibilities we applied a more stringent filter to the parental versus LM2 class comparison and also compared the LM0 populations to LM2. This resulted in a list of 42 candidate lung metastatic virulence genes (32 unique) generated by taking genes with either a six fold difference between parental and LM2 populations, or a three fold difference between LM0 and LM2. The metastagenicity and virulence gene lists were overlapping as some genes were associated with lung metastagenicity and had expression that was further increased in the LM2 populations. Nine biologically intriguing genes from either list were selected for functional validation. A final list of lung metastagenicity and virulence candidate genes was

generated by combining the 59 gene lung metastagenicity list with the nine genes that we selected for functional validation that were not already on the list for a total of 65 genes (54 unique). This list is shown in Supplementary Table 4.

Univariate and multivariate analysis of genes comprising the lung metastasis gene signature

In order to determine which of the 54 unique genes of the lung metastasis signature are associated with lung metastasis-free survival, we utilized a microarray dataset from 98 primary breast cancer patients treated at our institution. We excluded those with incomplete clinical annotations and/or if there was less than three years of clinical follow-up, resulting in 82 analyzable samples. At the time of tumor resection, these patients had an average age of 55.8 years (SD = 13.5 yrs), average tumor size of 3.68 cm (SD = 1.77 cm), and an average of 3.5 positive axillary lymph nodes (SD = 5.98). The vast majority of these patients received adjuvant chemotherapy and/or hormonal therapy.

For univariate analysis, each of the 54 unique genes of the lung metastasis signature was related to lung metastasis-free survival based on the Cox proportional hazards regression model. This process was also repeated for bone metastasis-free survival. The results of this analysis are shown in Supplementary Table 5. For multivariate analysis, the method of Beer et al²⁶ was used. In a leave-one-out cross-validated (LOOCV) manner, all 54 unique genes were used to generate a risk index for lung metastasis. In each round, using only the training cases, this risk index was defined as a linear combination of gene

expression values weighted by their estimated Cox model regression coefficients. The risk of the single training case was then determined. If the risk index for the training case was in the top 20th percentile of the risk index scores, then it was termed high-risk. Otherwise, it was termed low-risk. The 20th percentile was used as a cut-off because about 20 percent of the cases were expected to eventually develop lung metastasis.

Weighting each gene by its estimated Cox model coefficient for lung metastasis is a way to test the ability of the 54 genes to predict clinical high risk groups. A complementary approach is to test the ability of the genes to predict a biological group similar to the LM2 cell lines to see if this group is at high risk for developing lung metastasis. These two methods may not necessarily give the same results because each gene is weighted differently. For example, if many genes that better distinguish LM2 from the parental cell lines are not clinically meaningful, the two classifiers could give different results. To classify each of 82 samples in the MSKCC cohort into those that either resembled the LM2 cell lines or the parental MDA-MB-231 cell lines, a compound covariate classifier (BRBArray Tools 3.2) was used. Class membership into these two groups was determined by using the 54 gene lung metastasis signature. A compound covariate value was defined as a linear combination of gene expression values weighted by a t-statistic derived from comparing the LM2 cell lines (4173, 4175, and 4180) with two different passages of the parental MDA-MB-231 cell line (ATCC). The classification threshold was set as the midpoint of the sum of the mean values of

the compound covariate for each sample in the LM2 class or the ATCC class. Each of the 82 MSKCC samples was then predicted to be in the LM2 class if its compound covariate was closer to the LM2 class value or to be in the ATCC class if closer to the ATCC class value. Class survival analysis for lung metastasis-free survival and bone metastasis-free survival for the two classes of patients was then performed using the log-rank test.

Clustering of primary breast tumor data

For several reasons, using each gene of our lung metastasis signature in a linear combination as mentioned above, may have limitations in an analysis for a metastasis gene signature. One reason is because different tumors in a high risk group may have different combinations of individual genes. Furthermore, an experimentally-derived signature will likely contain features that are peculiar to the experimental system. In our case, we were hypothesizing that some of the genes in the experimental lung metastasis signature were serving as rare metastatic virulence genes, making it unlikely that they would be expressed by a bulk primary tumor population. Thus, to analyze the extent to which expression of the lung metastasis signature was similar to the LM2 cell lines, we applied unsupervised clustering methods using both the MSKCC data set and a second data set (Rosetta) comprised of 78 primary tumors⁹. The Rosetta data set utilized the Rosetta microarray platform. We were able to map 48 Affymetrix probe sets from our 54 unique genes to this platform. One of the 78 Rosetta

samples (sample 54) was omitted from the analysis because of a high number of missing values for many of our genes of interest.

Using BRBArray Tools 3.2, we performed hierarchical clustering to search for subgroups of patients that express the lung metastasis genes in a manner similar to the LM2 cell lines. A cluster reproducibility index R was used to evaluate the robustness of the clusters³⁵. The R measure is based on perturbing the expression data with Gaussian noise, re-clustering, and measuring the similarity of the new clusters to the original clusters. For each pair of samples in a cluster of the original data, the R measure is the proportion of the time they stay in the same cluster after perturbation and re-clustering over all pairs of samples, perturbations, and re-clustering. Clusters with high R value were identified and manual inspection appeared to reveal a group of primary breast cancers with concordant expression of many of the lung metastasis genes.

We also wanted to relate expression of the lung metastasis signature to the Rosetta poor-prognosis gene-expression signature⁹, estrogen receptor (ER) status, progesterone receptor (PR) status, HER2 status, and the basal/luminal breast cancer subtypes^{27,36}. Mapping of the 70 gene poor prognosis gene signature from the Rosetta platform to the Affymetrix platform resulted in 57 shared genes. ER and PR status was visualized using estrogen receptor alpha and progesterone receptor probes present on the Affymetrix U133A GeneChip or the Rosetta platform. HER2 status was determined by probes for ERBB2 and for GRB7³⁷. The probe for keratin 5 and keratin 17 were used as markers for the basal cell subtype and keratin 8 and keratin 18 for the luminal subtype²⁷. The

heatmap used to visualize gene expression was arranged so that the sample order was the same as determined by the hierarchical clustering results mentioned above.

Class prediction

From the MSKCC and the Rosetta data sets, it appeared that there exists a breast cancer subgroup of predominantly ER negative, poor prognosis, basal cell-like breast cancers that concordantly express many elements of the lung metastasis signature. Although useful for class discovery and analyzing relationships among clusters of genes, hierarchical clustering is not a statistical method for making class assignments. This is because partitioning samples into groups by inspection can be arbitrary and it does not provide a useful class predictor for new cases. However, recent work has described using class prediction methods for cancer subgroups defined by unsupervised clustering across data sets³⁸. Thus, we took advantage of the observation that a partial lung metastasis signature is expressed in two independent data sets.

The Rosetta data set was used as the training set to define the class labels used for prediction. We wished to identify two classes – samples that either did or did not express the lung metastasis signature in manner resembling the LM2 cell lines. Normalized data was imported into TIGR MultiExperiment Viewer 3.0.3 (ref. ³⁹) and the genes were median centered. The method of K-means was used to partition the training set based on the 48 genes of the lung metastasis signature shared by both microarray platforms. Choosing the right number of

clusters for K-means clustering is not obvious and is a long-standing problem. We estimated the K value based on a figure of merit⁴⁰, which assesses the predictive power of clustering using a left-out sample. This showed that a cluster number up to four resulted in a sharp decline in the figure of merit (lower score is better) and cluster numbers greater than this tended to show a higher error. To control for variation in results due to random initializations of the K-means algorithm, we also used K-means support, which produces consensus K-means clusters after multiple runs³⁹. Thus, the initial cluster number was set to four with 50 runs per iteration, the threshold percentage of occurrence in the same cluster was set at 70%, and 2000 K-means iterations were performed. Under these conditions four consensus clusters were produced and 36 of the 77 samples were unassigned.

The expression of the lung metastasis signature for each of the four consensus clusters was then evaluated for similarity to the LM2 cell lines by calculating the Pearson correlation between the cluster centroids and the centroid for the LM2 cell lines (Supplementary Figure S7). The mean centered gene expression data for the LM2 cell lines (4173, 4175, 4180) and two different passages of the MDA-MD-231 parental cells was used to calculate the LM2 centroid. From this analysis, cluster 3 had a Pearson correlation of 0.19 while the other clusters (including the unassigned samples) were anti-correlated (Supplementary Figure S7). Thus, the 13 members of cluster 3 were defined as a robust subgroup of tumors expressing the lung metastasis signature and all other samples were labeled as not expressing this signature. Repeated analysis with different

parameters used in K-means clustering confirmed the robustness for membership into these classes.

Because the 78 sample Rosetta training set and the 98 sample MSKCC test set were on different microarray platforms, both data sets were z-score transformed⁴¹. This was accomplished by taking the log2 transformed expression value of each gene, subtracting the mean expression value of that gene, and dividing this difference by the standard deviation. Each z-score transformed data set was then imported into BRBArray Tools 3.2. To guard against peculiarities of different class prediction methods, we used multiple predictors including 1-nearest neighbor, nearest centroid, and support vector machine with linear kernel and default penalty costs. In leave-one-out cross-validation each class prediction method correctly classified 95-96% of the Rosetta samples. Each of the 82 analyzable samples in the MSKCC data set was then classified to predict which belonged to the lung metastasis signature class. Results for each of the three prediction methods were similar. We used the consensus results, i.e. two out of the three. Survival analysis for lung metastasis and bone metastasis-free survival was then calculated using the log-rank test.

In an alternative approach to training the classifiers, we directly compared the lung metastasis signature centroid for the LM2 cell lines with each of the samples in the Rosetta data set using a Pearson correlation. This resulted in a range of correlations from -0.33 to 0.33. We selected an 80th percentile threshold corresponding to a correlation of greater than 0.15. These 16 samples were then

used in training for class prediction. In LOOCV, the class prediction methods correctly classified 68-92%, with 1-nearest neighbor being the worst and support vector machine being the best. Results after classification of the MSKCC data set were comparable to the K-means based classifier.

Rosetta poor prognosis classification

We were able to map 54 of the 70 Rosetta poor prognosis signature genes to the Affymetrix U133A platform. To ensure that this reduction in gene number does not significantly reduce the prognostic performance of the full signature we repeated the analysis of van't Veer et al⁹ using only the 54 genes that are also present on the Affymetrix platform. Using all 70 genes, 3 out of 34 poor prognosis cases were misclassified and 11 out of 44 good prognosis cases were misclassified (this was one fewer misclassification than reported by van't Veer et al.). Using the reduced subset of 54 genes, 5 poor prognosis cases were misclassified and 11 good prognosis cases were misclassified. Thus, the reduction in the signature had little impact on the performance of the classifier.

Each of the 82 breast cancer primaries from the MSKCC data set were assigned as having either a good prognosis signature or a poor prognosis signature. The method used by van't Veer et al. used binary data based on 5 year metastasis-free survival. Fourteen of the 82 MSKCC cases did not have at least five-years of follow-up and had to be excluded. For the remaining 68 cases, the van't Veer analysis, including LOOCV, was performed on z-transformed Affymetrix data. Classification was based on correlation with the good prognosis signature. While

van't Veer used a threshold of about 0.3 (the value used was not explicitly stated in their methods), we used 0. The results were that 5 out of 22 (23%) poor prognosis cases were misclassified and 19 out of 46 (41%) good prognosis cases were misclassified. The success of this classification was unlikely to be due to chance ($p=0.001$ based on 1000 permutations). The remaining 14 cases were classified in a similar manner, except using the 68 with 5-year survival as a training set. In this way, all 82 were classified as good or bad prognosis.

Clinical annotations, gene lists, and results of class assignments and predictions are collated in a workbook supplied as supplementary information.

Supplementary Table 1: Cell populations used in metastasis assays.

Metastatic propensity to bone and lung for all in vivo selected and single cell-derived populations used in the study.

Cell Line	Lung Metastatic Activity	Bone Metastatic Activity
<i>Parental</i>	-/+	+
1833	-/+	+++
1834	+	+
3475	++	+
3481	++	+
2293	+	+
2295	+	+
4142	+++	+
4173	+++	+
4175	+++	+
4180	+++	+
SCP 2	-/+	+++
SCP 3	+	+
SCP 6	-	-/+
SCP 21	-	-
SCP 25	-/+	++
SCP 26	-	-
SCP 28	+	+++
SCP 32	+	+
SCP 43	+	+
SCP 46	-/+	+++

Supplementary Table 2: Class comparison between parental MDA-MB-231 and LM2 cell lines selected to be highly metastatic to lung. Shown are 95 unique genes from 113 Affymetrix probe sets. Yellow marks 61 overexpressed probe sets and blue marks 52 underexpressed probe sets after a three-fold filter was applied.

Probe set	Fold Change	Gene Title	Gene Symbol
200665_s_at	407.01	secreted protein, acidic, cysteine-rich (osteonectin)	SPARC
203029_s_at	147.27	protein tyrosine phosphatase, receptor type, N polypeptide 2	PTPRN2
203030_s_at	97.07	protein tyrosine phosphatase, receptor type, N polypeptide 2	PTPRN2
207442_at	58.71	colony stimulating factor 3 (granulocyte)	CSF3
206172_at	48.52	interleukin 13 receptor, alpha 2	IL13RA2
206785_s_at	33.05	killer cell lectin-like receptor subfamily C, member 1 /// killer cell lectin-like receptor subfamily C, member 2	KLRC1 /// KLRC2
202310_s_at	20.03	collagen, type I, alpha 1	COL1A1
211534_x_at	15.67	protein tyrosine phosphatase, receptor type, N polypeptide 2	PTPRN2
221261_x_at	14.65	melanoma antigen, family D, 4 /// melanoma antigen, family D, 4	MAGED4
202947_s_at	13.50	glycophorin C (Gerbich blood group)	GYPC
204475_at	13.35	matrix metalloproteinase 1 (interstitial collagenase)	MMP1
217388_s_at	12.82	kynureninase (L-kynurenine hydrolase)	KYNU
205767_at	8.99	Epiregulin	EREG
201645_at	7.43	tenascin C (hexabrachion)	TNC
204698_at	6.77	Interferon stimulated gene 20kDa	ISG20
205623_at	6.75	Aldehyde dehydrogenase 3 family, member A1	ALDH3A1
212091_s_at	6.35	collagen, type VI, alpha 1	COL6A1
213711_at	6.34	keratin, hair, basic, 1	KRTHB1
210663_s_at	6.29	kynureninase (L-kynurenine hydrolase)	KYNU
204748_at	6.23	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	PTGS2
201720_s_at	5.83	Lysosomal-associated multispinning membrane protein-5	LAPTM5
203571_s_at	5.74	chromosome 10 open reading frame 116, adipose specific 2	C10ORF116
204205_at	5.29	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G	APOBEC3G
205463_s_at	5.02	platelet-derived growth factor alpha polypeptide	PDGFA
213194_at	4.86	roundabout, axon guidance receptor, homolog 1 (Drosophila)	ROBO1
212190_at	4.63	serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2	SERPINE2
220217_x_at	4.56	SPANX family, member C	SPANXC
221009_s_at	4.56	angiopoietin-like 4	ANGPTL4
201564_s_at	4.55	fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)	FSCN1
216268_s_at	4.47	jagged 1 (Alagille syndrome)	JAG1
201417_at	4.45	SRY (sex determining region Y)-box 4	SOX4
220922_s_at	4.40	SPANX family, member B1 /// SPANX family, member C	SPANXB1

			/// SPANXC
201288_at	4.26	Rho GDP dissociation inhibitor (GDI) beta	ARHGDIB
213428_s_at	4.24	collagen, type VI, alpha 1	COL6A1
220921_at	4.21	SPANX family, member B1	SPANXB1
33304_at	4.16	Interferon stimulated gene 20kDa	ISG20
205174_s_at	4.01	glutaminyl-peptide cyclotransferase (glutaminyl cyclase)	QPCT
210933_s_at	3.99	fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)	FSCN1
204470_at	3.89	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	CXCL1
201069_at	3.85	matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	MMP2
205399_at	3.76	doublecortin and CaM kinase-like 1	DCAMKL1
201061_s_at	3.71	Stomatin	STOM
221902_at	3.62	G protein-coupled receptor 153	GPR153
221760_at	3.59	mannosidase, alpha, class 1A, member 1	MAN1A1
219563_at	3.57	chromosome 14 open reading frame 139	C14orf139
211368_s_at	3.54	caspase 1, apoptosis-related cysteine protease (interleukin 1, beta, convertase)	CASP1
209030_s_at	3.42	immunoglobulin superfamily, member 4	IGSF4
202728_s_at	3.41	latent transforming growth factor beta binding protein 1	LTBP1
204385_at	3.24	kynureninase (L-kynurenine hydrolase)	KYNU
209505_at	3.24	nuclear receptor subfamily 2, group F, member 1	NR2F1
201325_s_at	3.21	epithelial membrane protein 1	EMP1
201721_s_at	3.21	Lysosomal-associated multispinning membrane protein-5	LAPTM5
206097_at	3.17	solute carrier family 22 (organic cation transporter), member 1-like antisense	SLC22A1LS
201324_at	3.15	epithelial membrane protein 1	EMP1
203417_at	3.12	microfibrillar-associated protein 2	MFAP2
208937_s_at	3.10	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein	ID1
219911_s_at	3.10	solute carrier organic anion transporter family, member 4A1	SLCO4A1
222182_s_at	3.07	CCR4-NOT transcription complex, subunit 2	CNOT2
222103_at	3.07	Activating transcription factor 1	ATF1
203585_at	3.06	zinc finger protein 185 (LIM domain)	ZNF185
221911_at	3.02	hypothetical protein LOC221810	LOC221810
216488_s_at	0.33	ATPase, Class VI, type 11A	ATP11A
205017_s_at	0.33	muscleblind-like 2 (Drosophila)	MBNL2
210046_s_at	0.33	isocitrate dehydrogenase 2 (NADP+), mitochondrial	IDH2
213075_at	0.33	olfactomedin-like 2A	OLFML2A
202149_at	0.32	neural precursor cell expressed, developmentally down-regulated 9	NEDD9
202610_s_at	0.32	cofactor required for Sp1 transcriptional activation, subunit 2, 150kDa	CRSP2
210340_s_at	0.32	colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)	CSF2RA
221011_s_at	0.32	likely ortholog of mouse limb-bud and heart gene /// likely ortholog of mouse limb-bud and heart gene	LBH
219959_at	0.31	molybdenum cofactor sulfurase	MOCOS
213537_at	0.31	major histocompatibility complex, class II, DP alpha 1	HLA-DPA1

202237_at	0.30	nicotinamide N-methyltransferase	NNMT
206473_at	0.30	membrane-bound transcription factor protease, site 2	MBTPS2
201428_at	0.30	claudin 4	CLDN4
201843_s_at	0.30	EGF-containing fibulin-like extracellular matrix protein 1	EFEMP1
202017_at	0.30	epoxide hydrolase 1, microsomal (xenobiotic)	EPHX1
202688_at	0.30	tumor necrosis factor (ligand) superfamily, member 10	TNFSF10
205018_s_at	0.29	muscleblind-like 2 (Drosophila)	MBNL2
203387_s_at	0.29	TBC1 domain family, member 4	TBC1D4
212372_at	0.28	myosin, heavy polypeptide 10, non-muscle	MYH10
205805_s_at	0.27	receptor tyrosine kinase-like orphan receptor 1	ROR1
216060_s_at	0.27	dishevelled associated activator of morphogenesis 1	DAAM1
203974_at	0.26	haloacid dehalogenase-like hydrolase domain containing 1A	HDHD1A
204149_s_at	0.25	glutathione S-transferase M4	GSTM4
210136_at	0.25	LOC388483	---
214040_s_at	0.24	gelsolin (amyloidosis, Finnish type)	GSN
213067_at	0.24	myosin, heavy polypeptide 10, non-muscle	MYH10
207379_at	0.24	EGF-like repeats and discoidin I-like domains 3	EDIL3
201137_s_at	0.23	major histocompatibility complex, class II, DP beta 1	HLA-DPB1
208306_x_at	0.23	major histocompatibility complex, class II, DR beta 3	HLA-DRB3
215193_x_at	0.23	major histocompatibility complex, class II, DR beta 3	HLA-DRB3
202986_at	0.23	aryl-hydrocarbon receptor nuclear translocator 2	ARNT2
206814_at	0.22	nerve growth factor, beta polypeptide	NGFB
204070_at	0.21	retinoic acid receptor responder (tazarotene induced) 3	RARRES3
202238_s_at	0.21	nicotinamide N-methyltransferase	NNMT
201842_s_at	0.21	EGF-containing fibulin-like extracellular matrix protein 1	EFEMP1
207620_s_at	0.18	calcium/calmodulin-dependent serine protein kinase (MAGUK family)	CASK
211990_at	0.18	Major histocompatibility complex, class II, DP alpha 1	---
202350_s_at	0.17	matrilin 2	MATN2
211907_s_at	0.16	par-6 partitioning defective 6 homolog beta (C. elegans) /// par-6 partitioning defective 6 homolog beta (C. elegans)	PARD6B
207214_at	0.16	serine protease inhibitor, Kazal type 4	SPINK4
211839_s_at	0.16	colony stimulating factor 1 (macrophage)	CSF1
208209_s_at	0.16	complement component 4 binding protein, beta	C4BPB
202145_at	0.14	lymphocyte antigen 6 complex, locus E	LY6E
211991_s_at	0.13	major histocompatibility complex, class II, DP alpha 1	HLA-DPA1
204238_s_at	0.12	chromosome 6 open reading frame 108	C6orf108
209394_at	0.10	acetylserotonin O-methyltransferase-like	ASMTL
208161_s_at	0.09	ATP-binding cassette, sub-family C (CFTR/MRP), member 3	ABCC3
209201_x_at	0.08	chemokine (C-X-C motif) receptor 4	CXCR4
210140_at	0.07	cystatin F (leukocystatin)	CST7
212942_s_at	0.07	KIAA1199	KIAA1199
217028_at	0.06	chemokine (C-X-C motif) receptor 4	CXCR4
214827_at	0.04	par-6 partitioning defective 6 homolog beta (C. elegans)	PARD6B

Supplementary Table 3: Overlapping genes between lung and bone metastasis signatures. The 113 probe sets (95 unique genes) from Supplementary Table 2 were overlapped with the 127 probe sets (102 unique genes) previously identified as the gene-expression signature of MDA-MB-231 cell populations that are highly metastatic to bone. Shown are 9 intersecting genes (11 probe sets) and whether each is up-regulated or down-regulated in either the bone metastasis signature or the lung metastasis signature.

Probe set	Description	Gene symbol	Bone	Lung
201417_at	SRY (sex determining region Y)-box 4	SOX4	down	up
203571_s_at	adipose specific 2	C10orf116	down	up
208161_s_at	ATP-binding cassette, sub-family C (CFTR/MRP), 3	ABCC3	down	down
211991_s_at	major histocompatibility complex, class II, DP alpha 1	HLA-DPA1	down	down
219563_at	chromosome 14 open reading frame 139	C14orf139	up	up
204475_at	matrix metalloproteinase 1 (interstitial collagenase)	MMP1	up	up
209201_x_at	Chemokine (C-X-C motif) receptor 4	CXCR4	up	down
220921_at	sperm protein associated with the nucleus, X chromosome, family member A1	SPANXA1	up	up
220922_s_at	sperm protein associated with the nucleus, X chromosome, family member A1	SPANXA1	up	up
215193_x_at	major histocompatibility complex, class II, DR beta 1	HLA-DRB1	down	down
201137_s_at	major histocompatibility complex, class II, DP beta 1	HLA-DPB1	down	down

Supplementary Table 4: Lung metastasis candidate genes. Shown are 54 unique genes from 65 Affymetrix probe sets representing genes associated with lung metastagenicity and virulence. Overexpressed fold change (yellow) and underexpressed fold change (blue) from comparing parental MDA-MB-231 and the LM2 cell lines are indicated.

Probe set	Fold Change	Gene Title	Gene Symbol
200665_s_at 212667_at	407.01	secreted protein, acidic, cysteine-rich (osteonectin)	SPARC
206172_at	48.52	interleukin 13 receptor, alpha 2	IL13RA2
206785_s_at	33.05	killer cell lectin-like receptor subfamily C, member 1 /// killer cell lectin-like receptor subfamily C, member 2	KLRC1 /// KLRC2
204475_at	13.35	matrix metalloproteinase 1 (interstitial collagenase)	MMP1
217388_s_at 210663_s_at	12.82	kynureninase (L-kynurenine hydrolase)	KYNU
205767_at	8.99	Epiregulin	EREG
201645_at	7.43	tenascin C (hexabrachion)	TNC
204698_at	6.77	interferon stimulated gene 20kDa	ISG20
205623_at	6.75	aldehyde dehydrogenase 3 family, member A1	ALDH3A1
213711_at	6.34	keratin, hair, basic, 1	KRTHB1
204748_at	6.23	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	PTGS2
201720_s_at 201721_s_at	5.83	Lysosomal-associated multispinning membrane protein-5	LAPTM5
203571_s_at	5.74	chromosome 10 open reading frame 116, adipose specific 2	C10orf116
213194_at	4.86	roundabout, axon guidance receptor, homolog 1 (Drosophila)	ROBO1
220217_x_at	4.56	SPANX family, member C	SPANXC
221009_s_at	4.56	angiopoietin-like 4	ANGPTL4
201564_s_at 210933_s_at	4.55	fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)	FSCN1
201417_at 201416_at	4.45	SRY (sex determining region Y)-box 4	SOX4
220922_s_at 220921_at	4.40	SPANX family, member B1 /// SPANX family, member C	SPANXB1 /// SPANXC
213428_s_at	4.24	collagen, type VI, alpha 1	COL6A1
204470_at	3.89	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	CXCL1
201069_at	3.85	matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	MMP2
201061_s_at	3.71	Stomatin	STOM
221902_at	3.62	G protein-coupled receptor 153	GPR153
221760_at	3.59	mannosidase, alpha, class 1A, member 1	MAN1A1
219563_at	3.57	chromosome 14 open reading frame 139	C14orf139
211368_s_at	3.54	caspase 1, apoptosis-related cysteine protease (interleukin 1, beta, convertase)	CASP1
209030_s_at	3.42	immunoglobulin superfamily, member 4	IGSF4

202728_s_at	3.41	latent transforming growth factor beta binding protein 1	LTBP1
209505_at	3.24	nuclear receptor subfamily 2, group F, member 1	NR2F1
201325_s_at	3.21	epithelial membrane protein 1	EMP1
201324_at		inhibitor of DNA binding 1, dominant negative helix-loop-helix protein	ID1
208937_s_at	3.10	CCR4-NOT transcription complex, subunit 2	CNOT2
222182_s_at	3.07	vascular cell adhesion molecule 1	VCAM1
203868_s_at	2.17	olfactomedin-like 2A	OLFML2A
213075_at	0.33	neural precursor cell expressed, developmentally down-regulated 9	NEDD9
202149_at	0.32	colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)	CSF2RA
210340_s_at	0.32	molybdenum cofactor sulfurase	MOCOS
219959_at	0.31	epoxide hydrolase 1, microsomal (xenobiotic)	EPHX1
202017_at	0.30	muscleblind-like 2 (Drosophila)	MBNL2
205018_s_at	0.29	LOC388483	---
205017_s_at		gelsolin (amyloidosis, Finnish type)	GSN
210136_at	0.25	myosin, heavy polypeptide 10, non-muscle	MYH10
214040_s_at	0.24	aryl-hydrocarbon receptor nuclear translocator 2	ARNT2
213067_at	0.24	retinoic acid receptor responder (tazarotene induced) 3	RARRES3
202986_at	0.23	EGF-containing fibulin-like extracellular matrix protein 1	EFEMP1
204070_at	0.21	matrilin 2	MATN2
201842_s_at	0.17	lymphocyte antigen 6 complex, locus E	LY6E
201843_s_at		major histocompatibility complex, class II, DP alpha 1	HLA-DPA1
202350_s_at	0.14	acetylserotonin O-methyltransferase-like	ASMTL
202145_at	0.13	ATP-binding cassette, sub-family C (CFTR/MRP), member 3	ABCC3
211991_s_at	0.10	KIAA1199	KIAA1199
213537_at	0.09	chemokine (C-X-C motif) receptor 4	CXCR4
209394_at	0.07	par-6 partitioning defective 6 homolog beta (C. elegans)	PARD6B
208161_s_at	0.06		
212942_s_at	0.04		
217028_at			
209201_x_at			
214827_at			

Supplementary Table 5. Expression of Genes in the Lung Metastasis Signature Correlated to Lung Metastasis-Free Survival in Breast Cancer Patients. A Cox proportional hazards model was used to relate gene expression changes of the 54 gene lung metastasis signature to lung metastasis-free survival in 82 breast cancer patients.

Probe set	Gene Symbol	Hazard Ratio	Lower 95%	Upper 95%	p-value
204070_at	RARRES3	0.434	0.291	0.648	0.00001
221009_s_at	ANGPTL4	2.991	1.661	5.388	0.00005
203571_s_at	C10orf116	0.608	0.467	0.792	0.00047
202728_s_at	LTBP1	3.364	1.467	7.711	0.00074
205017_s_at	MBNL2	3.133	1.357	7.231	0.00169
201564_s_at	FSCN1	1.975	1.280	3.047	0.00201
201324_at	EMP1	2.997	1.411	6.369	0.00272
210340_s_at	CSF2RA	1.805	1.212	2.687	0.00283
204475_at	MMP1	1.313	1.064	1.619	0.00742
212942_s_at	KIAA1199	1.617	1.076	2.431	0.02083
204470_at	CXCL1	1.356	1.076	1.708	0.02191
204748_at	PTGS2	1.451	1.030	2.043	0.02628
202986_at	ARNT2	0.746	0.542	1.026	0.06494
213067_at	MYH10	0.674	0.429	1.060	0.06899
213075_at	OLFML2A	0.434	0.165	1.139	0.07305
222182_s_at	CNOT2	0.365	0.120	1.108	0.07775
206785_s_at	KLRC1	0.752	0.544	1.040	0.08261
208161_s_at	ABCC3	0.776	0.574	1.048	0.10283
202145_at	LY6E	0.704	0.437	1.136	0.13893
202017_at	EPHX1	0.678	0.387	1.186	0.17169
209505_at	NR2F1	0.806	0.579	1.121	0.21238
210663_s_at	KYNU	1.235	0.887	1.718	0.21883
210136_at	MBP	1.431	0.809	2.532	0.22674
219959_at	MOCOS	1.359	0.830	2.226	0.23861
201061_s_at	STOM	0.613	0.267	1.408	0.24098
213428_s_at	COL6A1	1.542	0.722	3.293	0.25386
219563_at	C14orf139	0.657	0.319	1.355	0.25881
220217_x_at	SPANXC	0.773	0.474	1.261	0.28465
213537_at	HLA-DPA1	0.786	0.493	1.253	0.33430
213711_at	KRTHB1	1.100	0.899	1.347	0.36209
201645_at	TNC	1.195	0.805	1.772	0.37407
201721_s_at	LAPTM5	1.305	0.634	2.687	0.48354
201842_s_at	EFEMP1	0.865	0.570	1.313	0.49742
213194_at	ROBO1	1.216	0.699	2.113	0.49865
214040_s_at	GSN	1.167	0.717	1.901	0.51734
220921_at	SPANXB1	0.892	0.612	1.301	0.54461

209030_s_at	IGSF4	0.755	0.300	1.899	0.54672
202350_s_at	MATN2	0.907	0.658	1.252	0.55728
208937_s_at	ID1	1.156	0.716	1.866	0.56958
209394_at	ASMTL	0.816	0.400	1.667	0.58735
221760_at	MAN1A1	0.890	0.522	1.519	0.66920
205767_at	EREG	1.058	0.814	1.374	0.67603
206172_at	IL13RA2	1.061	0.691	1.629	0.78848
211368_s_at	CASP1	1.065	0.663	1.710	0.79193
201069_at	MMP2	1.079	0.592	1.966	0.80346
203868_s_at	VCAM1	1.065	0.576	1.969	0.83993
204698_at	ISG20	0.973	0.743	1.273	0.84223
205623_at	ALDH3A1	0.957	0.598	1.531	0.85511
201416_at	SOX4	0.941	0.462	1.913	0.86571
214827_at	PARD6B	0.972	0.648	1.458	0.88897
217028_at	CXCR4	0.953	0.482	1.884	0.88906
221902_at	GPR153	0.964	0.524	1.773	0.90587
212667_at	SPARC	0.969	0.489	1.922	0.92818
202149_at	NEDD9	1.033	0.510	2.092	0.92853

Supplementary Table 6. Lung Metastasis Signature Genes Used to Classify Primary Breast Cancers Expressing the Lung Metastasis Signature. All genes from Table 1 are shown.

p-value	UG cluster	Gene symbol	Description
<0.000001	Hs.118400	FSCN1	Fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)
<0.000001	Hs.83169	MMP1	Matrix metalloproteinase 1 (interstitial collagenase)
<0.000001	Hs.9613	ANGPTL4	Angiopoietin-like 4
0.000006	Hs.74120	C10orf116	Chromosome 10 open reading frame 116
0.00002	Hs.789	CXCL1	Chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
0.000355	Hs.196384	PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
0.000444	Hs.185568	KRTHB1	Keratin, hair, basic, 1
0.000506	Hs.109225	VCAM1	Vascular cell adhesion molecule 1
0.000627	Hs.17466	RARRES3	Retinoic acid receptor responder (tazarotene induced) 3
0.001263	Hs.368256	LTBP1	Latent transforming growth factor beta binding protein 1
0.004365	Hs.444471	KYNU	Kynureninase (L-kynurenine hydrolase)
0.005179	Hs.421986	CXCR4	Chemokine (C-X-C motif) receptor 4
0.006426	Hs.77667	LY6E	Lymphocyte antigen 6 complex, locus E
0.007153	Hs.410900	ID1	Inhibitor of DNA binding 1, dominant negative helix-loop-helix protein
0.010871	Hs.255149	MAN1A1	Mannosidase, alpha, class 1A, member 1
0.032361	Hs.388589	NEDD9	Neural precursor cell expressed, developmentally down-regulated 9
0.03713	Hs.115263	EREG	Epiregulin
0.046859	Hs.98998	TNC	Tenascin C (hexabrachion)
0.053773	Hs.357901	SOX4	SRY (sex determining region Y)-box 4
0.05492	Hs.157986	MOCOS	Molybdenum cofactor sulfurase
0.062067	Hs.165725	CNOT2	CCR4-NOT transcription complex, subunit 2
0.071707	Hs.436200	LAPTM5	Lysosomal-associated multispinning membrane protein-5
0.079271	Hs.153647	MATN2	Matrilin 2
0.080391	Hs.156682	IGSF4	Immunoglobulin superfamily, member 4
0.096189	Hs.306692	EMP1	Epithelial membrane protein 1
0.097858	Hs.105434	ISG20	Interferon stimulated gene 20kDa
0.119096	Hs.280311	MYH10	Myosin, heavy polypeptide 10, non-muscle
0.124785	Hs.301198	ROBO1	Roundabout, axon guidance receptor, homolog 1 (Drosophila)
0.213167	Hs.361748	NR2F1	Nuclear receptor subfamily 2, group F, member 1
0.230817	Hs.125715	MBNL2	Muscleblind-like 2 (Drosophila)
0.25087	Hs.367877	MMP2	MMP2
0.254227	Hs.446537	GSN	Gelsolin (amyloidosis, Finnish type)
0.255766	Hs.531581	GPR153	G protein-coupled receptor 153

0.274128	Hs.336046	IL13RA2	Interleukin 13 receptor, alpha 2
0.345846	Hs.357004	OLFML2A	Olfactomedin-like 2A
0.36839	Hs.6111	ARNT2	Aryl-hydrocarbon receptor nuclear translocator 2
0.423864	Hs.111779	SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)
0.507582	Hs.2490	CASP1	Caspase 1, apoptosis-related cysteine protease (interleukin 1, beta, convertase)
0.650845	Hs.76224	EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1
0.75516	Hs.520937	CSF2RA	Colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)
0.764736	Hs.439776	STOM	Stomatin
0.830009	Hs.512576	KLRC1	Killer cell lectin-like receptor subfamily C, member 1
0.830451	Hs.415997	COL6A1	Collagen, type VI, alpha 1
0.843369	Hs.458420	ASMTL	Acetylserotonin O-methyltransferase-like
0.846476	Hs.575	ALDH3A1	Aldehyde dehydrogenase 3 family, memberA1
0.867387	Hs.89649	EPHX1	Epoxide hydrolase 1, microsomal (xenobiotic)
0.899238	Hs.90786	ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
0.926966	Hs.914	HLA-DPA1	Major histocompatibility complex, class II, DP alpha 1

Supplementary Figure Legends

Supplementary Figure S1. Single cell-derived progenies (SCPs) of MDA-MB-231 cells have a uniform Rosetta-type poor prognosis gene signature and variation in gene expression correlating with metastatic behavior.

Fifty-four of the 70 Rosetta poor prognosis genes were present on the Affymetrix U133A microarray platform and performed comparably to the original 70 genes in predicting five year metastasis-free survival (Supplementary Methods). The 54 genes successfully classified patients in the MSKCC cohort with at least five years of clinical follow-up (77% correct classification of poor prognosis and 59% correct classification of good prognosis, $p=0.001$) and was used to assign all 82 patients into good versus poor prognosis groups. **a**, The gene expression centroid for the MSKCC good prognosis and poor prognosis groups are shown at the top of the heatmap. Below this is the expression of each of the 54 shared Rosetta poor prognosis genes for the SCPs. For presentation purposes, the intensity of the good prognosis and poor prognosis centroids was increased by a factor of six to more closely match the overall intensity of the cell line data. The gene expression data is median centered with yellow being up-regulated and blue being down-regulated. Genes overexpressed (red bar) and underexpressed (green bar) in poor prognosis tumors are shown on the bottom. **b**, Uniformity in the expression of a Rosetta-type poor prognosis signature is shown using a pairwise Pearson correlation comparing this signature among the SCPs and indicated MDA-MB-231 cell lines. ATCC refers to parental MDA-MB-231 cells. **c**, Variation in gene expression among SCPs is represented in three dimensions using multi-dimensional scaling and reveals three distinct groups with similarities in gene expression. **d**, Bioluminescence imaging (BLI) of representative SCPs from each of the three groups taken 7 weeks after tailvein or intracardiac xenografting.

Supplementary Figure S2. Confirmation of protein expression for lung metastasis signature genes used in functional validation. The indicated MDA-MB-231 in vivo selected populations were analyzed by **a**, Western blotting for SPARC and ID1, **b**, ELISA for MMP1 and MMP2, or by **c**, flow cytometry analysis for VCAM1 and IL13R α 2 staining.

Supplementary Figure S3. Validation of combination transgenic parental MDA-MB-231 cell lines transduced with lung metastasis genes. Parental MDA-MB-231 cells were retrovirally transduced. Northern blot analysis identifies exogenous transcripts for **a**, *SPARC*, *ID1*, and *MMP1*, **b**, *VCAM1*, *IL13R α 2*, and *MMP2*, or **c**, *CXCL1*, *EREG*, and *COX2*. These genes were expressed either individually (which is shown for *SPARC*, *ID1*, and *MMP1*), or in combinations of three or six. Puro represents the empty vector control.

Supplementary Figure S4. Parental MDA-MB-231 cells overexpressing lung metastasis genes are not enhanced in bone metastatic activity. Parental MDA-MB-231 cells retrovirally transduced with vector controls or various combinations of lung metastasis genes, and highly bone metastatic 1833 cells were injected into the left cardiac ventricle of immunocompromised mice. Bioluminescent imaging was used to monitor the development of bone metastases. Representative mice from cohorts of 5 animals each were used for presentation purposes.

Supplementary Figure S5. Lung metastasis signature genes are able to distinguish patients at high risk for developing lung but not bone metastasis. Patients in the MSKCC cohort were classified using a linear combination of each of the 54 lung metastasis signature genes. **a**, Each gene was weighted by its estimated Cox model regression coefficient for either lung or bone metastasis to classify patients into a clinical low-risk group (blue) or a high-risk group (red and brown). **b**, Each of the 54 genes was weighted by a t-statistic

derived from comparing its expression between LM2 cell lines with the parental MDA-MB-231 cell lines to classify patients as being more similar to either the parental cell lines (blue) or the LM2 cell lines (red and brown). Shown are survival curves for lung metastasis-free survival (top) and bone metastasis-free survival (bottom) with p-values.

Supplementary Figure S6. Identification of a subgroup of primary breast cancers that express the lung metastasis signature in the Rosetta data set.

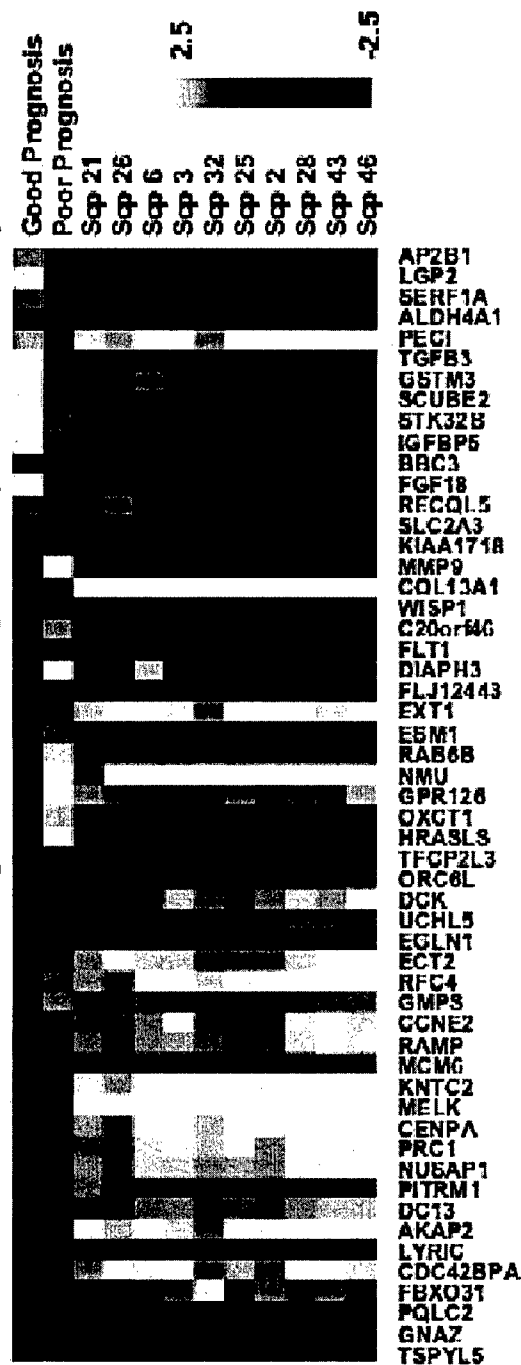
Hierarchical clustering of primary breast carcinomas from a cohort of 77 breast cancer patients⁹ was performed using 48 lung metastasis candidate genes that mapped to the Rosetta microarray⁹. A dendrogram resulting from clustering of the tumors is shown at the top, with tumors from patients that developed metastasis denoted by black circles. The rows corresponding to the nine lung metastasis genes that were functionally validated in mice are shown in greater detail (*middle panel*) with the names of each gene on the right. The Rosetta poor-prognosis signature for each of these tumors is displayed with genes that are overexpressed (*red bar*) and underexpressed (*green bar*) in poor prognosis tumors indicated on the left. Expression of HER2, estrogen receptor/progesterone receptor status, and basal and luminal keratins is also shown²⁷. The gene expression data is centered with red/gold indicating up-regulation and green/blue indicating down-regulation. A sub-cluster with a cluster reproducibility index of 0.81 (*dashed red box*) groups tumors that tended to express the lung metastasis signature in a manner resembling the LM2 cell lines.

Supplementary Figure S7. Classification of primary breast cancers that express the lung metastasis signature used in class prediction training. K-means support clustering was used to partition the breast primaries from the Rosetta data set into four clusters (see Supplementary methods section). Shown are the lung metastasis gene-expression signature centroids for each of four

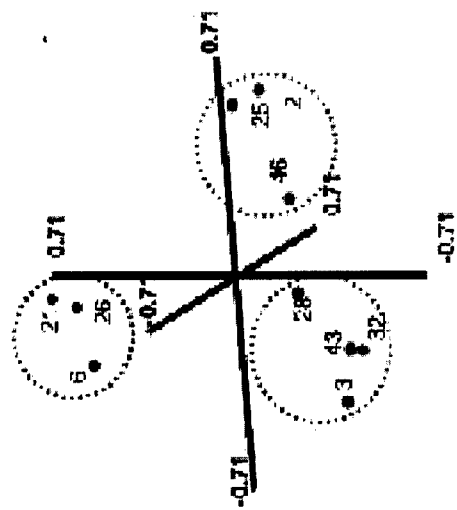
consensus clusters. Cluster 0 refers to patients that were unassigned to any of the four clusters. Also shown are the centroids for the LM2 cell lines (4173, 4175, 4180) and two different passages of the parental MDA-MB-231 cell line (ATCC). Similarity of each consensus cluster to the LM2 cell line is visualized by hierarchical clustering and the Pearson correlation values are shown in the table below the heatmap. The names of the 48 lung metastasis signature genes that mapped to the Rosetta microarray platform are shown on the right, with the genes that were functionally validated shown in red. Yellow represents up-regulated genes, and blue represents down-regulated genes. Members of cluster 3 were defined as a robust subgroup of tumors expressing the lung metastasis signature and all other samples were labeled as not expressing this signature. These class labels were used to train a classifier.

a

Rosetta Poor Prognosis Gene Signature (van't Veer et al.)



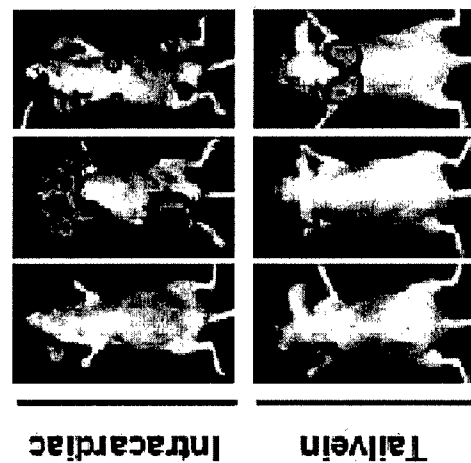
c

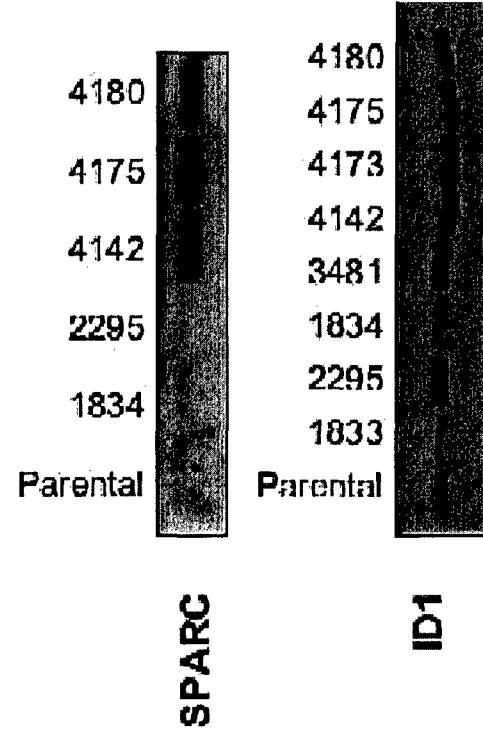


b

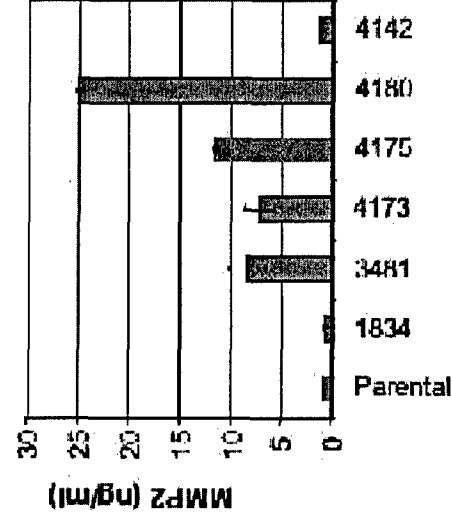
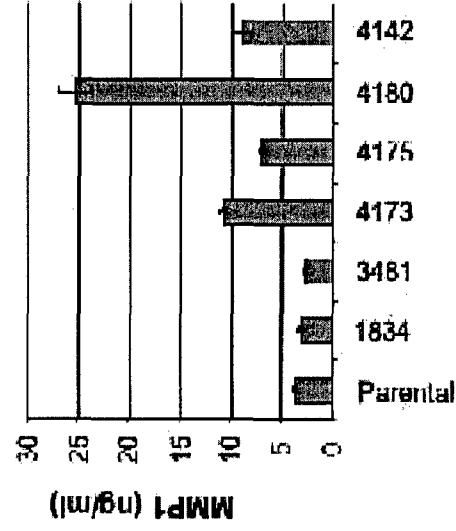
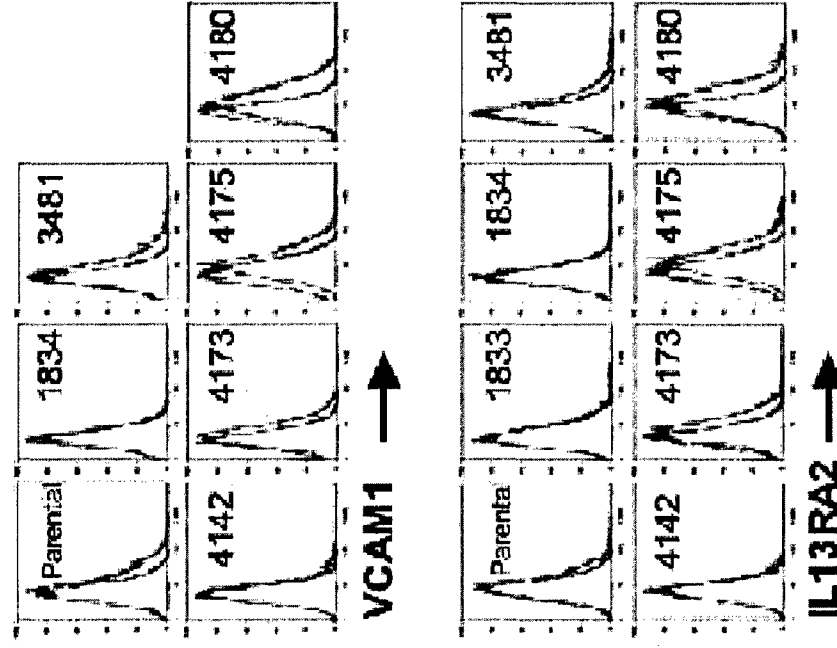
	ATC	Scp 2	Scp 3	Scp 6	Scp 21	Scp 25	Scp 26	Scp 28	Scp 32	Scp 43	Scp 46	4173	4175	4180
ATOC	100	0.95	0.92	0.94	0.95	0.95	0.96	0.96	0.95	0.96	0.93	0.97	0.95	0.97
Scp 2	0.95	100	0.95	0.95	0.90	0.96	0.91	0.98	0.96	0.98	0.96	0.95	0.95	0.94
Scp 3	0.92	0.95	100	0.95	0.88	0.95	0.93	0.95	0.97	0.96	0.96	0.93	0.94	0.95
Scp 6	0.94	0.95	0.93	100	0.92	0.94	0.93	0.94	0.94	0.96	0.95	0.93	0.95	0.93
Scp 21	0.95	0.90	0.88	0.92	100	0.92	0.94	0.96	0.91	0.90	0.91	0.92	0.91	0.92
Scp 25	0.95	0.96	0.95	0.94	0.92	100	0.94	0.96	0.97	0.97	0.96	0.94	0.96	0.94
Scp 26	0.96	0.91	0.93	0.93	0.94	0.94	100	0.94	0.93	0.93	0.93	0.94	0.93	0.95
Scp 28	0.96	0.93	0.95	0.94	0.93	0.96	0.94	100	0.97	0.96	0.96	0.96	0.96	0.94
Scp 32	0.95	0.96	0.97	0.94	0.91	0.97	0.93	0.97	100	0.97	0.96	0.96	0.97	0.94
Scp 43	0.96	0.93	0.96	0.96	0.90	0.97	0.93	0.96	0.97	100	0.96	0.97	0.97	0.95
Scp 46	0.93	0.96	0.96	0.95	0.91	0.96	0.93	0.96	0.96	0.96	100	0.93	0.94	0.94
4173	0.97	0.95	0.93	0.93	0.92	0.94	0.94	0.96	0.96	0.97	0.93	100	0.93	0.97
4175	0.96	0.96	0.94	0.95	0.91	0.96	0.93	0.96	0.97	0.97	0.94	0.93	100	0.96
4180	0.97	0.94	0.93	0.93	0.92	0.94	0.95	0.94	0.94	0.95	0.94	0.97	0.96	100

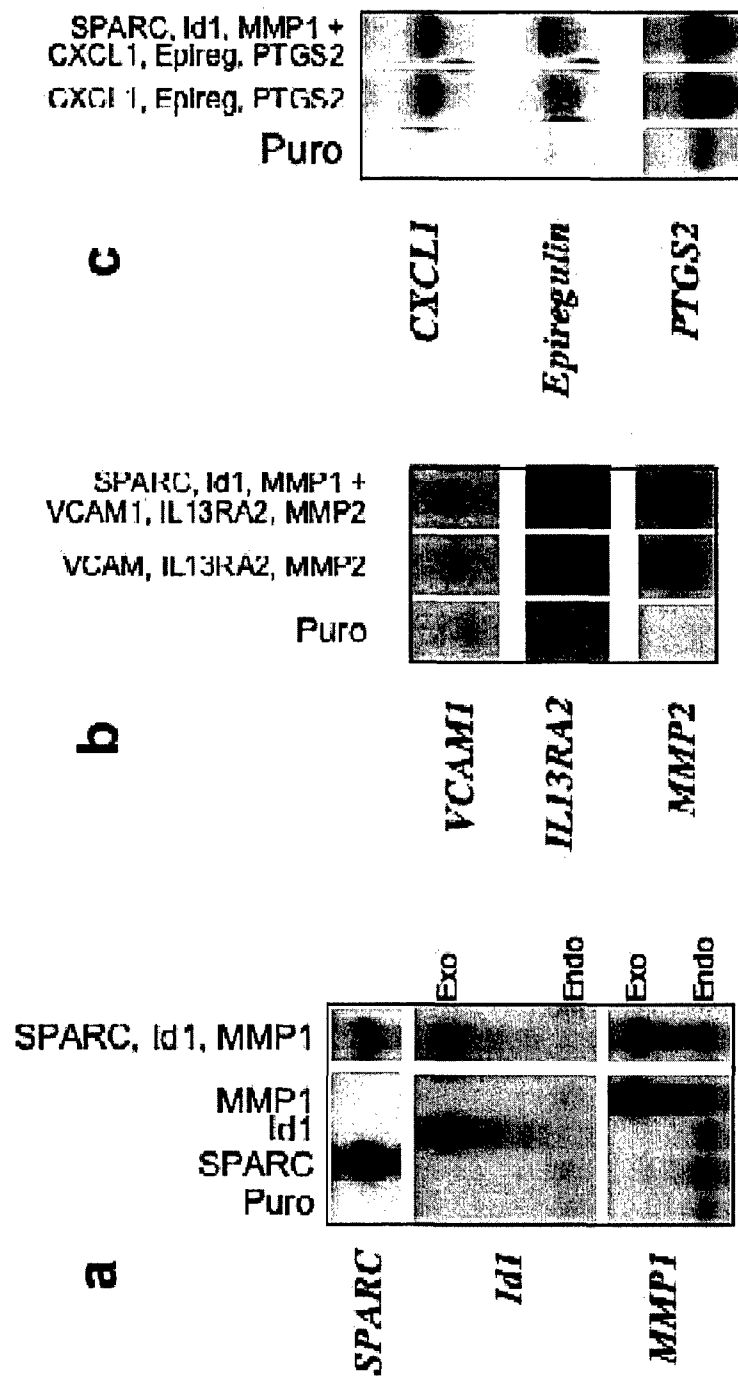
SCP6 SCP2 SCP32

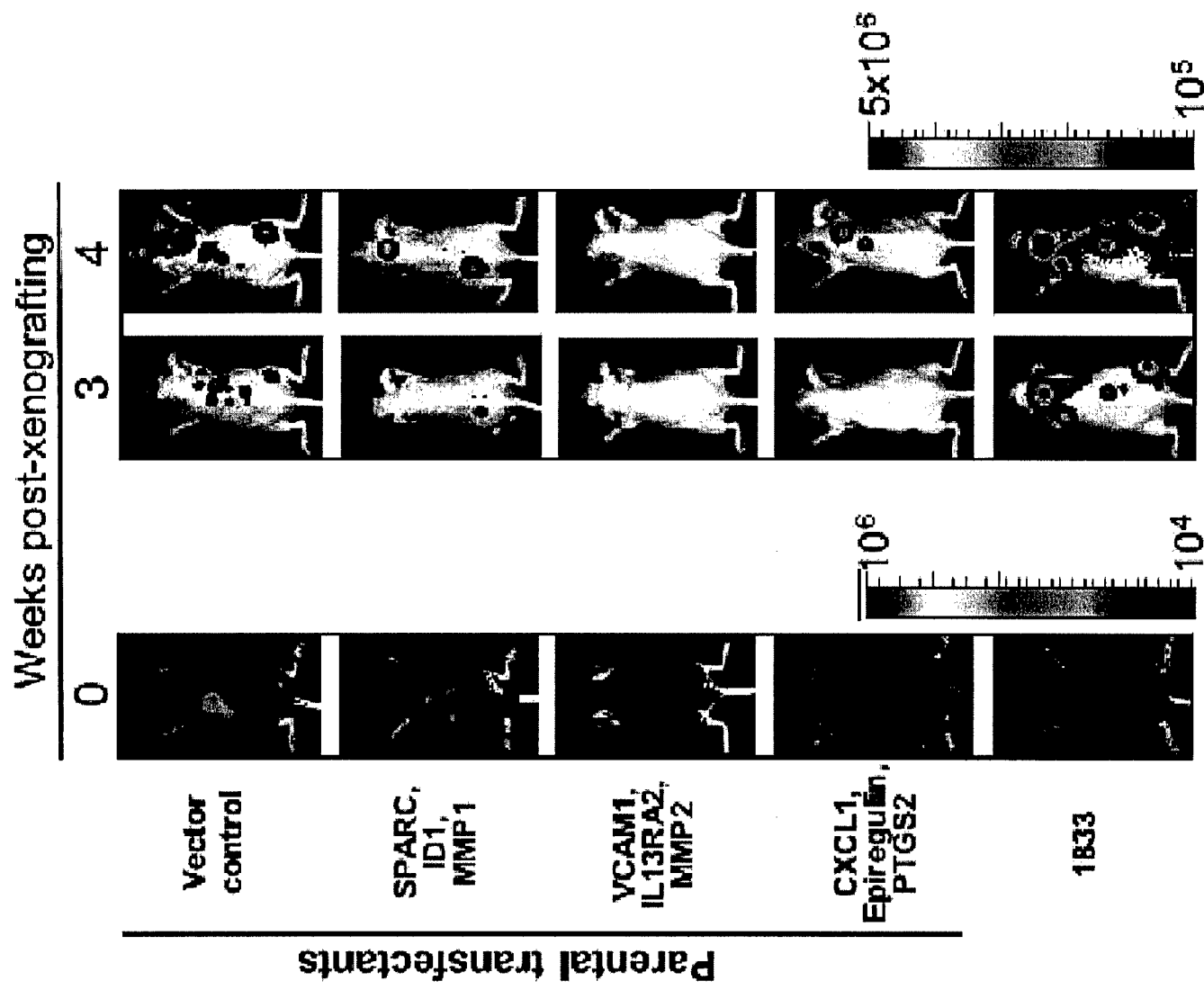




c

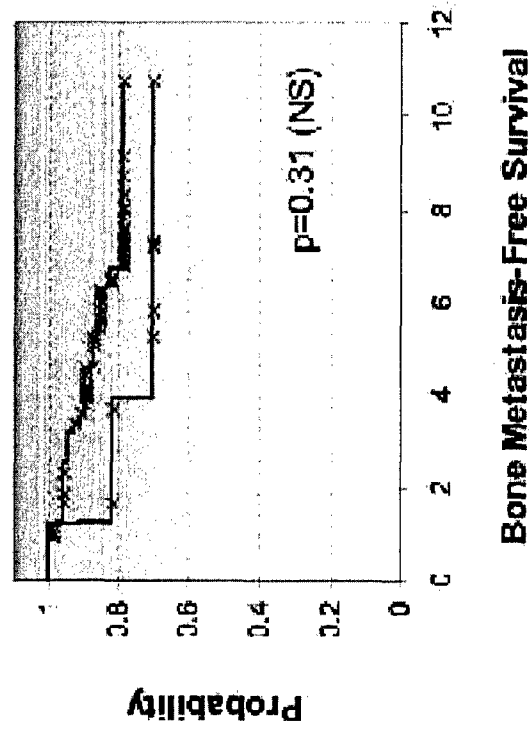
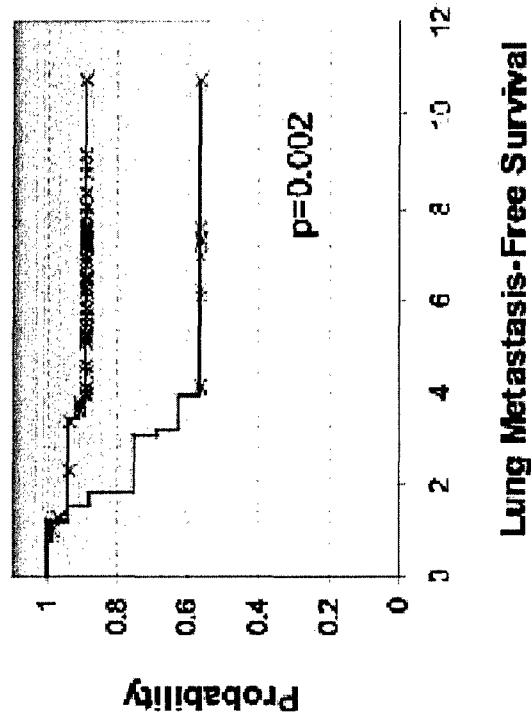






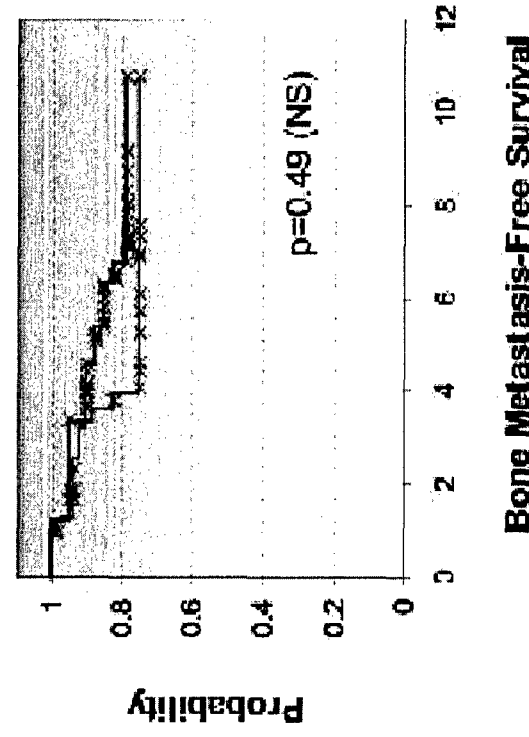
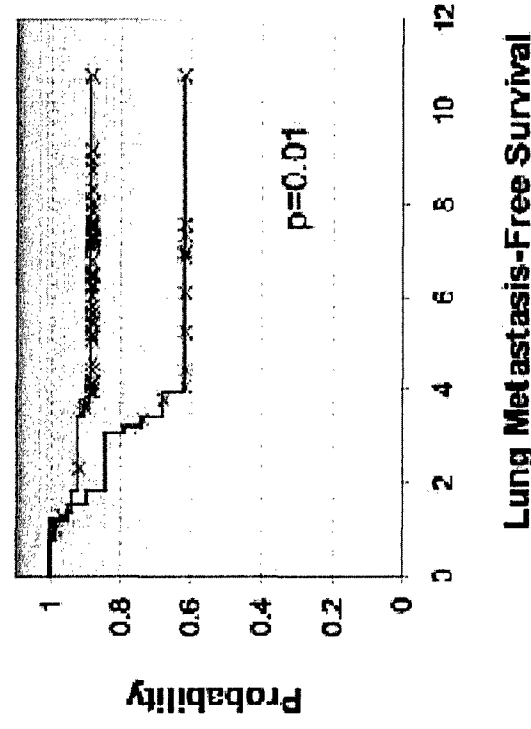
a

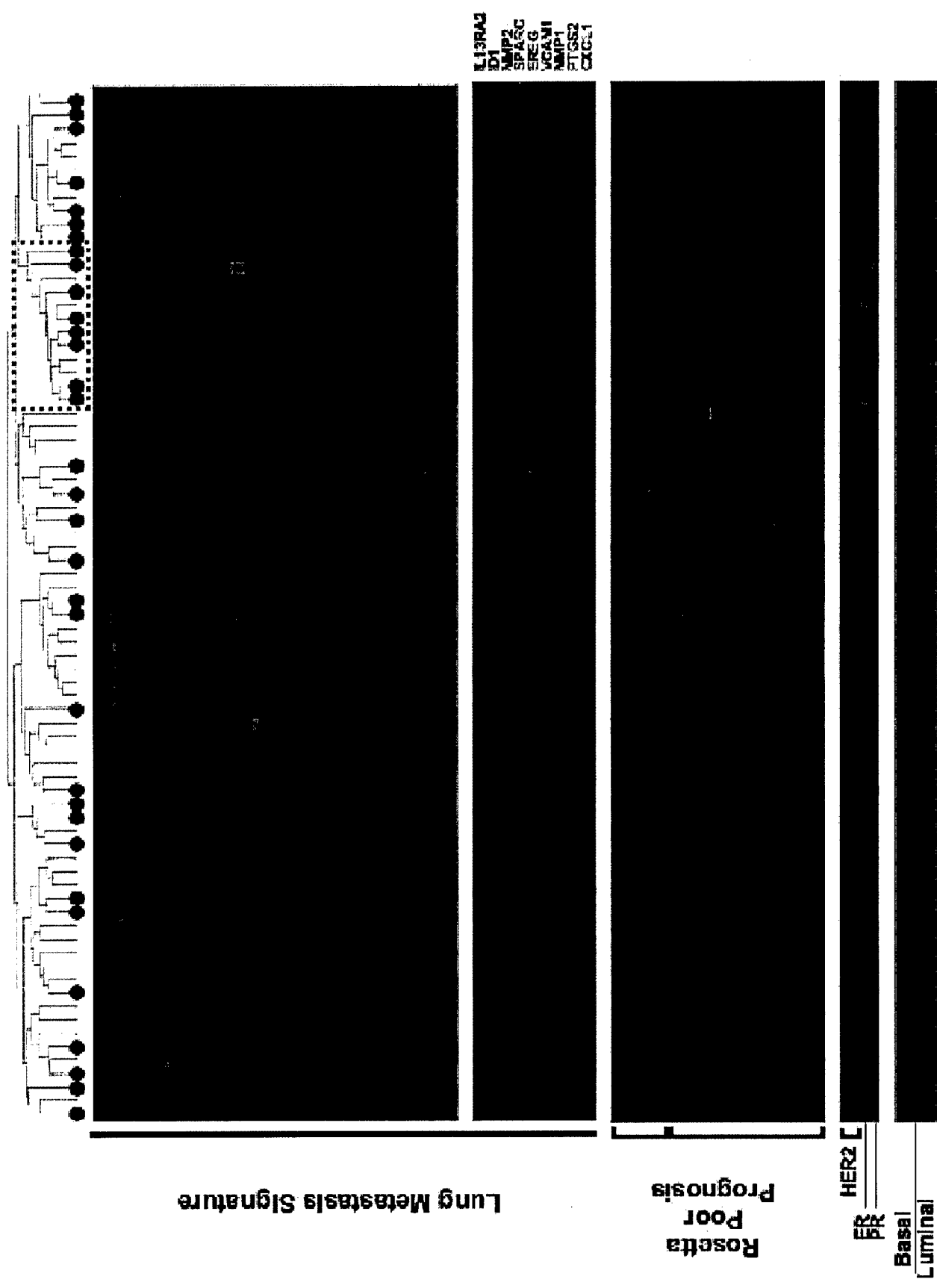
Weighted by Clinical Hazard Ratios



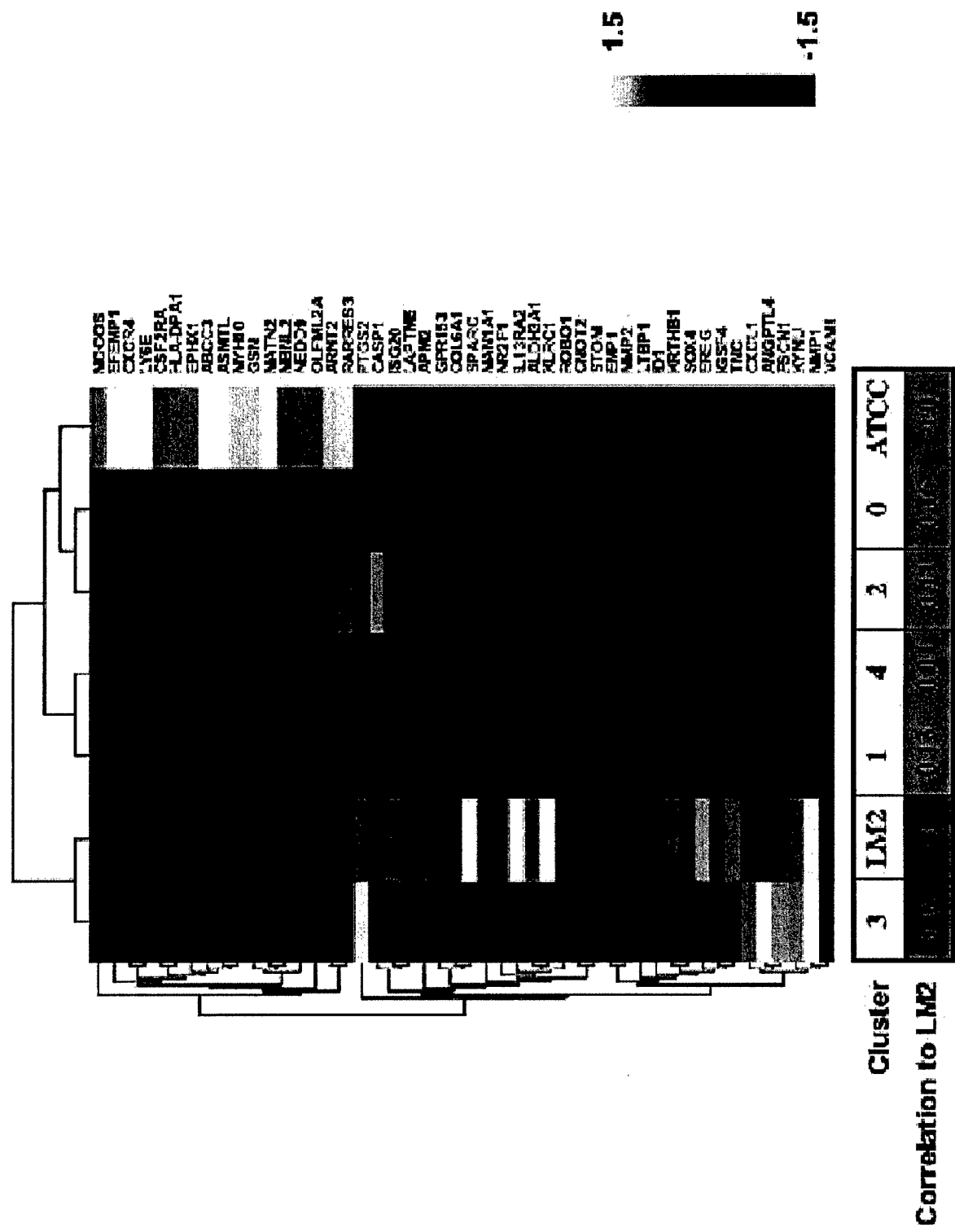
b

Weighted by ATCC vs LM2 T-Statistic





Supplementary Figure S6



Cluster					
3	LM2	1	4	2	0 ATCC
Correlation to LM2					
0.0	0.0	0.0	0.0	0.0	0.0

Supplementary Figure S7

LIST OF PERSONNEL

Brogi E	Co investigator
Danso M	Clinical Fellow
Doane A	Graduate assistant
Donaton M	Post doctoral Fellow
Gerald W	Principle investigator
Giri D	Clinical Fellow
Hudis C	Co investigator
Lal P	Clinical Fellow
Olshen A	Co investigator
Panageas K	Co investigator
Tan L	Co investigator
VanZee K	Co investigator
Zhang L	Post doctoral Fellow

BIBLIOGRAPHY

Bhargava R, Gerald W, Lal P, and Chen B. Epidermal growth factor receptor (EGFR) gene amplification in breast cancer: Correlation with mRNA and protein expression and absence of common activating mutations. *Mod Pathol*. Epub ahead of press 2005.

Dechow TN, Pedranzini L, Leitch A, Leslie K, Gerald WL, Linkov I, Bromberg JF. Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C. *Proc Natl Acad Sci U S A*. 2004 101(29):10602-7.

Doane A, Danso M, Lal P, Donaton M, Zhang L, and Gerald W. Estrogen Receptor-Negative Breast Cancer with an Active Hormone Response Pathway: Therapeutic Implications. Abstract presentation AACR 2005

Doane A, Danso M, Lal P, Donaton M, Zhang L, Hudis C, and Gerald W. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and proliferative response to androgen. Submitted

Donaton M, Giri D, Olshen A, Panageas K, Levcovici S, Lal P, Brogi E, Hudis C, VanZee K, Tan L, Gerald W *Comprehensive gene expression analysis of paired primary breast carcinomas and lymph node metastases*. Abstract presentation American Association of Cancer Research, 2003.

Giri D, Donaton M, Olshen A, Panageas K, Levcovici S, Lal P, Brogi E, Hudis C, VanZee K, Tan L, Gerald W. *Gene expression differences between paired primary and metastatic breast carcinomas*. Abstract presentation United States and Canadian Academy of Pathology, 2003.

Kang Y., He W, Gupta G, Tulley W, Serganova I, Chen C., Manova-Todorova K, Blasberg R, Gerald W and Massagué J. The Smad4 Tumor Suppressor Mediates Pro-Metastatic TGF β Gene Responses in Breast Cancer Bone Metastasis. Submitted

Lal P, Donaton M, Giri D, Chen B, Gerald W Molecular Diagnosis of Breast Cancer Therapeutic Biomarkers Using Oligonucleotide Arrays Abstract presentation USCAP 2005.

Minn A, Kang Y, Serganova I, Gupta G, Giri D, Doubrovin M, Ponomarev V, Gerald W, Blasberg R, Massague J. Distinct organ-specific metastasis potential of individual breast cancer cells and primary tumors. *J Clin Invest.* 115: 44-55, 2005

Minn A, Gupta G, Siegel P, Bos P, Shu W, Giri D, Viale A, Olshen A, Gerald W, Massague J. Genes that predict and mediate breast cancer metastasis to the lung. In press *Nature*.